Chemical Similarity based on Graph Edit Distance: Efficient Implementation and the Challenges of Evaluation

Roger Sayle, John May, Noel O’Boyle, Andrew Grant, Stefan Senger and Darren Green

NextMove Software, Cambridge, UK
2D CHEMICAL SIMILARITY

• Fingerprints
  – Feature-based (MACCS keys)
  – Path-based (Daylight, ChemAxon, OpenBabel)
  – Radial (ECFPn, FCFP, RDKit Morgan)

• Maximum Common Subgraph
  – Connected MCES (cMCS)
  – Disconnected MCES (dMCS)
  – Td MCS (Ed Duesbury)

• Others (Lingos, Descriptor Vectors, etc.)
LIMITATIONS OF FINGERPRINTS

• Similarity based on small “local” substructures.
• Saturation of features/Chemical Space.
  – Many peptides/proteins/nucleic acids have identical FPs.
  – For alkanes, C16 should be more similar to C18 than C20.
  – Identical FPs in Chemistry Toolkit Rosetta benchmark.
  – PubChem “similar compounds” uses 90% threshold.
• No distinction between atom type changes.
  – Chlorine to Bromine more conservative than to Boron.
  – Tautomers often have low similarity
• Stereochemistry is poorly handled.
  – Either not represented or isomers have low similarity.
FP COUNTER INTUITION

Daylight Tanimoto: 0.44

Daylight Tanimoto: 0.39
The Genesis of SmallWorld

- Algorithmically, the way molecules are compared for 2D similarity hasn’t changed much in decades: FPs.
- Compared to 3D methods using Linux clusters and GPUs, 2D fingerprints are seen as good-enough.
- Andrew Grant questioned whether, given the known deficiencies with binary fingerprints, it is possible to do better with the hardware currently available?
THE OLD EMPEROR HAD NO CLOTHES


**Briem & Lessel Benchmark**

80.67%  79.38%  75.96%  75.60%  66.17%  65.23%  64.00%

- SmallWorld
- ECFP_4
- FCFP_8
- LINGOs
- MACCS
- ChemAxon
- Daylight

7th Sheffield Conference on Chemoinformatics, Sheffield, UK  5th July 2016
THE SEARCH FOR CHEMICAL SIMILARITY

- Bemis-Murcko Scaffold 224/554 41.18%
- Match Molecular Pairs 68/79 86.08%
- Molecular Formula 40/48 83.33%
- R-groups (anti-scaffold) 358/712 50.28%
- Scaffold with attachments 106/208 50.96%
- Anonymous topology 52/57 91.23%
- Element graph 16/17 94.12%
In 1965, Vladimir Levenshtein introduced string edit distance as similarity measure between strings.

The minimum number of insertions, deletions and substitutions to transform one string to another.

Thanks to efficient dynamic programming algorithms of Needleman-Wunsch and Smith-Waterman, sequence alignment is at the core of bioinformatics.

Prior to string edit distance, similarity between sequences was performed by heuristic methods locally comparing short words, or $k$-tuples.
GRAPH EDIT DISTANCE

• Graph Edit Distance (GED) is the extension of this concept to graphs, as the minimum number of edit operations required to transform one graph into another.

• Edit operations consist of insertions, deletions and substitutions of nodes and edges (atoms and bonds).
EXAMPLE EDIT OPERATIONS

Benzene

Chlorobenzene

Benzoxazole

Pyridine

Fluorobenzene

Benzothiazole
EXAMPLE EDIT OPERATIONS

Benzene

Cyclohexane

Thiazole

Tetrahydrofuran

Histidine

Histidine Zwitterion
EXAMPLE EDIT OPERATIONS

Ticlodipine

Clopidogrel

Penicillin G

Amoxicillin

7th Sheffield Conference on Chemoinformatics, Sheffield, UK 5th July 2016
EXAMPLE EDIT OPERATIONS

Sildenafil (Viagra)

Vardenafil (Levitra)

Sumatriptan (Imitrex)

Zolmitriptan (Zomig)
GED AND MCES

• Unfortunately, calculating GED has been shown to be a generalization of calculating Maximum Common Subgraph (MCS), a favorite of chemists but known to be computationally very expensive (NP-Hard).
### Chemical Edit Distance Search

**Query**

![Chemical Structure](image)

**SMILES**

```
OC(=O)C(=O)OC[c1ccccc1]c0c3cc(n2cc(=c(n1)C(=O)O)nc2)c2cc(c1)n1
```

**DataSet**

ChemBL 20

**Search Type**

SmallWorld

---

### Results

<table>
<thead>
<tr>
<th>Compound</th>
<th>Distance</th>
<th>ECFP4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEMBL1667753</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>CHEMBL2296002</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>CHEMBL25</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>CHEMBL499817</td>
<td>1</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 6,146 entries

- **Identical**
- **Hydrogen Substitution**
- **Hybridisation Change**
- **Minor Transmutation**
- **Major Transmutation**
- **Deletion**

(2.4 s Elapsed)
### SmallWorld Chemical Edit Distance Search

**Query**

![Chemical Structure](image)

- **SMILES**: `CC(=O)Oc1cccccc1C(=O)O`
- **DataSet**: CHEMBL 20
- **Advanced Type**: SmallWorld

### Advanced Options

- **Distance**: 2
- **Terminal**: 0
- **Ring**: 0
- **Linker**: 0
- **Mutation**: 0
- **Substitution**: 0
- **Hybridisation**: 0
- **Atom Type Match**: 0

### Results

<table>
<thead>
<tr>
<th>Compound</th>
<th>Distance</th>
<th>ECFP4</th>
<th>TDN</th>
<th>TUP</th>
<th>RDN</th>
<th>RUP</th>
<th>LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEMBL1978455</td>
<td>2</td>
<td>0.00</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHEMBL2009730</td>
<td>2</td>
<td>0.00</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHEMBL98018</td>
<td>2</td>
<td>0.08</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHEMBL98067</td>
<td>2</td>
<td>0.06</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 30,948 entries
Fortunately, recent developments in theoretical computer science and advances in processing power and disk storage finesse the problem.

THE SECRET... PRE-PROCESSING

• Imagine in advance enumerating, canonicalizing and storing all subgraphs for a given molecule, sorted by the number of bonds.

• The task of finding the largest common subgraph then becomes an almost trivial ordered intersection.

• Modern state-of-the-art cheminformatics systems can easily handle databases several orders of magnitude larger than the current largest non-virtual databases (i.e. many billions of subgraphs).
SMALL WORLD CHEMICAL SPACE
All traditional MCS algorithms start from one bond and grow; SmallWorld starts from N bonds and shrinks.

In typical databases, highly similar molecules which cause problems for MCS, are found quickly, typically terminating the search.

The work of Jean-Louis Raymond et al. on GDB-15, using Brendan McKay’s GENG to enumerate all possible graphs, significantly over estimates the number of (sub)graphs encountered in practice.
A GRAND UNIFIED THEORY

• The three fundamental forces of cheminformatics:
  – Molecular Identity
  – Substructure Search
  – Chemical Similarity

• Bemis and Murcko Scaffolds
• Shuffenhauer Scaffold Trees
• Feature, Path and Radial Fingerprints
• Matched Molecular Pairs
DIFFERENT EDGE TYPES
BEWARE OF DALKE WORMHOLES
EFFICIENT SUBGRAPH ENUMERATION

• A connected Maximum Common Edge Subgraph (MCES) with one less bond is formed by either (i) deleting a bond to a terminal atom, or (ii) deleting a ring (cyclic) bond.

• Partitioning cyclic from acyclic bonds can be done efficiently in O(N) time, and even this only needs to be recalculated after deleting a ring bond, as deleting terminal bonds doesn’t affect ring membership.
# Subgraph Counts of Molecules

<table>
<thead>
<tr>
<th>Name</th>
<th>Atoms</th>
<th>MW</th>
<th>Anon SGs</th>
<th>Elem SGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benzene</td>
<td>6</td>
<td>78</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Cubane</td>
<td>8</td>
<td>104</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>Ferrocene</td>
<td>11</td>
<td>186</td>
<td>3,154</td>
<td>3,219</td>
</tr>
<tr>
<td>Aspirin</td>
<td>13</td>
<td>180</td>
<td>127</td>
<td>332</td>
</tr>
<tr>
<td>Dodecahedrane</td>
<td>20</td>
<td>260</td>
<td>440,473</td>
<td>440,473</td>
</tr>
<tr>
<td>Ranitidine</td>
<td>21</td>
<td>314</td>
<td>436</td>
<td>1,207</td>
</tr>
<tr>
<td>Clopidrogel</td>
<td>21</td>
<td>322</td>
<td>10,071</td>
<td>22,170</td>
</tr>
<tr>
<td>Morphine</td>
<td>21</td>
<td>285</td>
<td>176,541</td>
<td>496,467</td>
</tr>
<tr>
<td>Amlodipine</td>
<td>28</td>
<td>409</td>
<td>58,139</td>
<td>147,128</td>
</tr>
<tr>
<td>Lisinopril</td>
<td>29</td>
<td>405</td>
<td>24,619</td>
<td>34,496</td>
</tr>
<tr>
<td>Gefitinib</td>
<td>31</td>
<td>447</td>
<td>190,901</td>
<td>337,174</td>
</tr>
<tr>
<td>Atorvastatin</td>
<td>41</td>
<td>559</td>
<td>3,638,523</td>
<td>6,019,427</td>
</tr>
</tbody>
</table>
### Molecule DB Size Distribution

<table>
<thead>
<tr>
<th>≤ Bond Count</th>
<th>% MDDR 2011.2</th>
<th>% NCBI PubChem</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 20 bonds</td>
<td>6%</td>
<td>14%</td>
</tr>
<tr>
<td>≤ 25 bonds</td>
<td>18%</td>
<td>30%</td>
</tr>
<tr>
<td>≤ 30 bonds</td>
<td>36%</td>
<td>55%</td>
</tr>
<tr>
<td>≤ 35 bonds</td>
<td>56%</td>
<td>77%</td>
</tr>
<tr>
<td>≤ 40 bonds</td>
<td>73%</td>
<td>89%</td>
</tr>
<tr>
<td>≤ 45 bonds</td>
<td>83%</td>
<td>93%</td>
</tr>
<tr>
<td>≤ 50 bonds</td>
<td>89%</td>
<td>95%</td>
</tr>
<tr>
<td>≤ 55 bonds</td>
<td>92%</td>
<td>97%</td>
</tr>
<tr>
<td>≤ 60 bonds</td>
<td>93%</td>
<td>98%</td>
</tr>
<tr>
<td>≤ 65 bonds</td>
<td>94%</td>
<td>98%</td>
</tr>
<tr>
<td>≤ 70 bonds</td>
<td>95%</td>
<td>99%</td>
</tr>
</tbody>
</table>
SmallWorld lattice: Bold circles denote indexed molecules, thin circles represent virtual subgraphs.
The solid circle denotes a query structure which may be either an indexed molecule or a virtual subgraph.
The first iteration of the search adds the neighbors of the query to the “search wavefront”.

7th Sheffield Conference on Chemoinformatics, Sheffield, UK 5th July 2016
Each subsequent iteration propagates the wavefront by considering the unvisited neighbors of the wavefront.
At each iteration, “hits” are reported as the set of indexed molecules that are members of the wavefront.
SMALLWORLD SEARCH

The search terminates once sufficient indexed neighbors have been found (or a suitable iteration limit is reached).
SMALLWORLD SEARCH
SMALLWORLD SEARCH
Current DB Statistics

• As of 20th June 2016, the SmallWorld index has
• 19,043,283,752 nodes (~19B or \(2^{34}\) nodes)
• 76,037,538,012 edges (~76B or \(2^{36}\) edges)
  – 21,312,527,888 terminal edges
  – 40,955,851,425 ring (deletion/closure) edges.
  – 13,769,158,699 linker (deletion/insertion) edges.
• Run-time storage requirements: ~2.85 Tbytes.
• Average degree (fan-out) of node: ~8
• 1.22B acyclic nodes, 2.78B have single ring.
HEATMAP OF SMALLWORLD UNIVERSE
CONNECTED VS. DISCONNECTED MCS

• A frequent question which maximum common subgraph methods is whether to perform connected or disconnected MCS.

• One solution to this dilemma with SmallWorld is to allow an additional “edit operation” that can eliminate a vertex of degree two.

• This captures the similarity between two molecules that differ in the length of a linker or by the size of a ring [classic challenges with MCS similarity].
A useful feature of graph edit distance is intuitive interpretation of $\text{GED}(A,B) < \text{GED}(A,C)$. 

Representing Activity Cliffs
Tricky cases: what doesn’t work

Tricky case from Briem & Lessel benchmark, found as top hit by ECFP6, but with 20 bond changes (3 rings).
CONCLUSIONS

• The sub-linear behaviour of SmallWorld’s nearest neighbor calculation 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 is likely to become the basis of most chemical similarity calculations within a decade.
ACKNOWLEDGEMENTS

• In memoriam Andy Grant, thank you for everything.
• AstraZeneca R&D, Alderley Park, U.K.
• GlaxoSmithKline, Stevenage, U.K.
• Relay Therapeutics, Boston, U.S.A.
• Jose Batista, OpenEye Scientific Software, Germany.
• Jameed Hussain, Chemical Computing Group, U.K.

• Thank you for your time, Any questions?
J. ANDREW GRANT (1963-2012)

Andy and I at OpenEye EuroCUP 2008

7th Sheffield Conference on Chemoinformatics, Sheffield, UK 5th July 2016
ALGORITHM REFERENCES


EXAMPLE SEARCH STATISTICS (2013)

• A Briem & Lessel benchmark run, searching for the (upto) 10 nearest neighbors of each of 380 drug-sized query molecules takes 8m42s (less than 1.4 seconds per query) on single thread i7-2600/16GB.

• The median distance for each search was 8 edits, with the shortest finishing after 2 edits and the furthest reaching 24 edits.

• Limiting to 10 edits reduced search time to 7 mins.

• Worst case “wavefront” size was 6,624,624 nodes which occurred at distance/iteration nine.

7th Sheffield Conference on Chemoinformatics, Sheffield, UK 5th July 2016