SMCM: A SIMPLE METRIC FOR MOLECULAR COMPLEXITY

Tharun Kumar Allu※and Tudor I. Oprea

※Department of Computer Science and Office of Biocomputing, University of New Mexico School of Medicine
MSC08 4560, 1 University of New Mexico, Albuquerque, NM 87131-5196, USA.
Emails: tharun@ender.unm.edu and toprea@salud.unm.edu

Abstract

Molecular complexity is a topic of interest in early lead discovery. We propose a simple metric for molecular complexity, based on OEChem. This metric takes into account the number of (vicinal) chiral centers, the number of (fused) rings, branching, geminal substitutions (among other factors).

Introduction

Lead discovery requires an increasingly large number of combinatorial libraries to be evaluated *in silico*. Among the parameters used to prioritize compounds for synthesis, reagent availability, the cost of starting materials and the possibility of identifying few-steps, high-yield products, have been the most important. Since molecular similarity/diversity metrics have been used to evaluate combinatorial libraries, we propose another parameter, aimed at evaluating **molecular complexity**. The need to evaluate the complexity of a molecule, in order to judge which molecules are easier to synthesize, has been discussed since the early days of combinatorial chemistry. However, molecular weight (combined with simple descriptors such as LogP or polarizability) have been the only ones applied by chemists. We derive a simple molecular complexity metric, **SMCM**, starting from the number of different atom types and bonds, the number of chiral centers and the number of (vicinal) chiral centers, the number of (different) geminal substitutions, the number of rings and fused ring systems in the molecule and the number of rings which are fused, and the number of *spiro* carbons.

Method

We used OEChem 1.0, a library for Chemistry and Chemical Informatics, to process SMILES as input, then compute the following:

- Atom Types (C, c, C=, C#, N, n, N^+/n^+, N=, N#, O, o, O^-, O=, S, s, S=, =S=, F, Cl, Br, I, H, B, P=, nH, Polar H)
- The number of single bonds, double Bonds, triple bonds per Atom Type
- The number of rings
- The number of fused ring systems.
- The number of chiral centers
- The number of vicinal chiral centers
- The number of geminal substituted carbons (e.g., -C(x1)(x2)- or C(x1)(x2)x3)
- Carbon
  - Spiro?
  - Chiral Center?
  - Aromatic?
  - Hybridization (sp, sp^2, sp^3)
  - Geminal?
  - Charged? (unlikely, but possible)
- Nitrogen
  - Charged?
  - Genuine charge? (e.g., exclude [N+]([O-])=O or N=[N+]=[N-])
  - Aromatic?
- Oxygen
  - Charged?
  - Aromatic? (e.g., furan)
- Sulfur
  - Aromatic?
  - Chiral Center? (e.g., x1-S(=O)-x2)
  - Number of double bonds
- Phosphorus
  - Chiral Center? (e.g., x1-P(=O)(x2)-x3)
  - Number of double bonds
- First four Halogens, Boron, Silicon, Hydrogen (polar, total)

SMCM counts all the above, plus the number of ring systems and the number of fused ring systems using a formula provided by Geoff Skillman (OpenEye):

Number of Fused rings in each #RingSystem = #bonds - #atoms + 1 - #spiro atoms.

We considered different weighting schemes in SMCM for each of the atoms present in the molecule, then computed SMCM as a function of the sum of weights of each feature observed in the molecule.

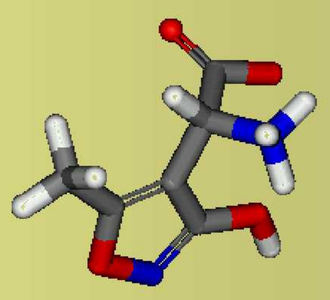
SMCM Weighting Schemes

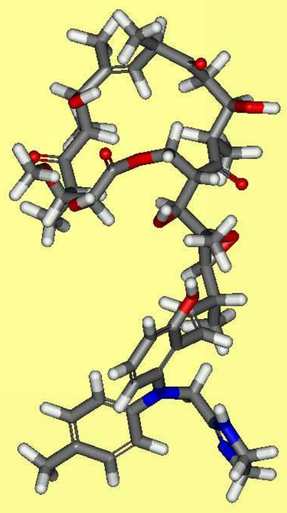
We tried several weighting schemes, in order to capture "complexity" in general. One SMCM scheme, which was more promising, is presented below.

Feature	*Weight*
F, Cl, Br, I, C	1 * no. Of each Atom
B, O, S, N	1.5 * no. Of each Atom
H	0.2 * no. Of each Atom
Polar H (attached to O,N,S)	1 * no. Of each Atom
S=, P=, C=, N=, O=	2 * no. Of each Atom
C#, N#	2 * no. Of each Atom
=S=	2 * no. Of each Atom
N+/n+, N-/n-, O-/o-	2.5 * no. Of each Atom
c,n,s,o	2 * no. Of each Atom
nH	2 * no. Of each Atom
SpiroCarbon	4 * no. Of each Atom
GeminalCarbon	3 * no. Of each Atom
ChiralCenters	$\log_2$(no. Of Chiral Centers)
AdjacentChiralCenters	$2^{(\text{no. Of each Atom})}$
No of Rings	$\log_2$ (no. Of Rings)
RingSystems	4 * no. Of RingSystems
TotalFusedRings	$2^{(\text{no. Of total Fused rings})}$
Freely Rotatable Bonds	-0.7 * no. Of each bond

Example 1:

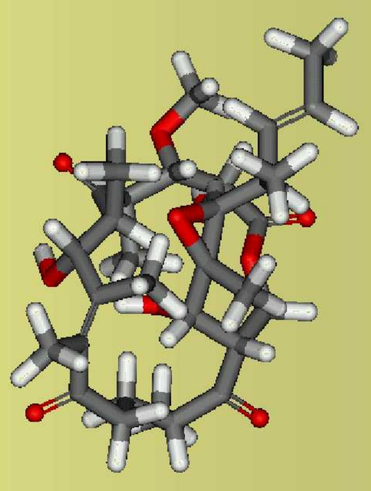
This is relatively simple molecule depending on all the metrics computed like 4 hydrogen atoms, 2 sp^3 carbon atoms, 1 sp^2 carbon with double bond, 1 nitrogen atom, 2 oxygen atoms, 1 oxygen with double bond, 3 aromatic carbon atoms, 1 aromatic nitrogen, 1 aromatic oxygen, 4 polar hydrogen atoms, 2 freely rotatable bonds, 1 ring system with 1 ring totally gives this molecule a complexity of 28.9





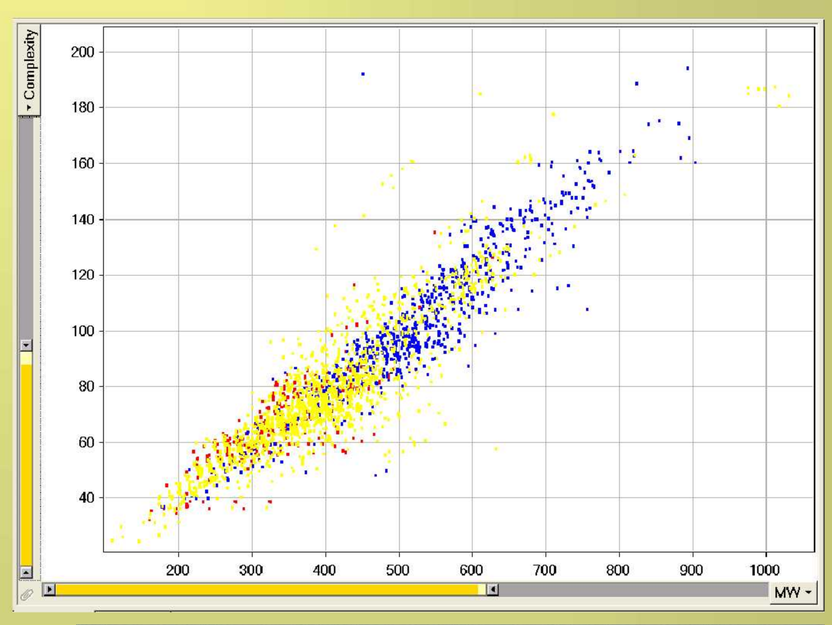
Example 2:

This is a more complex molecule with 59 hydrogen atoms, 27 sp^3 carbon atoms, 9 sp^2 carbon atoms, 2 nitrogen atoms, 1 nitrogen with double bond, 8 oxygen atoms, 4 oxygen atoms with double bonds, 12 aromatic carbon atoms, 6 polar hydrogen atoms, 10 freely rotatable bonds, 13 chiral centers, 12 vicinal chiral centers, 5 rings, 5 ring systems which contribute the complexity to be 4226.82



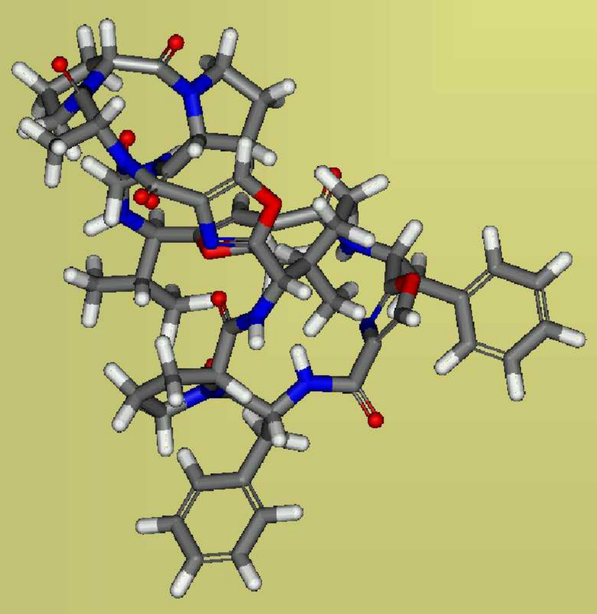
Example 3.

We get from the molecule 45 hydrogen atoms, 23 sp^3 carbon atoms, 8 sp^2 carbon atoms, 6 oxygen atoms, 4 oxygen atoms with double bonds, 3 polar hydrogen atoms, 5 freely rotatable bonds, 12 chiral centers, 12 vicinal chiral centers, 2 rings, 2 ring systems which computes up to 4173.08



SMCM vs. MW.

With the exception of vicinal chiral/substituted carbons, the increase in **SMCM** (feature counts) is paralleled by an increase in MW (R^2=0.846, N=3565). This could reflect our inability to perceive chirality correctly, and the lack of negative values for "easily accessible" synthetic blocks.



Example 4:

This example is interesting as it has many functional groups which we use for computing complexity. The metrics of this molecule are 67 hydrogen atoms, 29 sp^3 carbon atoms, 9 sp^2 carbon atoms, 9 nitrogen atoms, 9 oxygen atoms with double bonds, 22 aromatic carbon atoms, 2 aromatic nitrogen atoms, 3 aromatic oxygen atoms, 6 polar hydrogen atoms, 7 freely rotatable bonds, 7 chiral centers, 2 vicinal chiral centers, 9 rings, 3 ring systems, 7 fused rings, with alll these metrics using the weighing scheme we compute the complexity of the molecule to be 296.977

Summary and Outlook

SMCM is designed to assist medicinal and combinatorial chemists in evaluating virtual and existing libraries. Our evaluation on a set of 3565 biologically active molecules indicates that, in its current implementation, **SMCM** cannot replace molecular weight as a filter for complexity.

So far, **SMCM** does *not* evaluate "ease of synthesis" as seen by a chemist. We plan to introduce SMARTS-based negative values for fragments that are routinely used by chemists (e.g, sulfonamides, peptide bonds, esters), as well as for molecules that are routinely available from vendors in precursor form (e.g., sugars, quinazolines, penicillins). We will also improve chirality perception.

References:

1. OEChem Reference Manual.
2. Daylight Theory Manual.
3. M.M. Hann, A.R. Leach, G. Harper, J. Chem. Inf. Comput. Sci. 2001, 41, 856-864

The University of New Mexico ◆ Health Sciences Center
SCHOOL OF MEDICINE