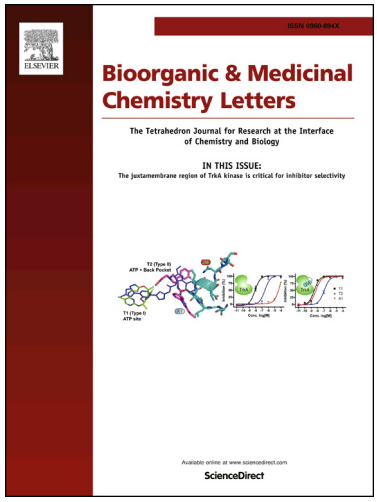
Accepted Manuscript



A Path Based Approach to Assessing Molecular Complexity

John R. Proudfoot

PII: S0960-894X(17)30243-3
DOI: http://dx.doi.org/10.1016/j.bmcl.2017.03.008
Reference: BMCL 24760

To appear in: *Bioorganic & Medicinal Chemistry Letters*

Received Date: 18 January 2017
Revised Date: 3 March 2017
Accepted Date: 4 March 2017

Please cite this article as: Proudfoot, J.R., A Path Based Approach to Assessing Molecular Complexity, *Bioorganic & Medicinal Chemistry Letters* (2017), doi: http://dx.doi.org/10.1016/j.bmcl.2017.03.008

ELSEVIER

Bioorganic & Medicinal Chemistry Letters
journal homepage: www.elsevier.com

A Path Based Approach to Assessing Molecular Complexity

John R. Proudfoot

Boehringer Ingeleim Pharmaceuticals Inc
900 Ridgebury Road, PO Box 368
Ridgefield CT06877, USA

E-mail address: john.proudfoot@discoverybytes.com

ARTICLE INFO

Article history:
Received
Revised
Accepted
Available online

Keywords:
Molecular Complexity
Atom Complexity
Shannon Entropy
Approved Drugs

ABSTRACT

An atom environment, path based approach to calculating molecular complexity is described. Based on Shannon's equation, the method transforms the number and diversity of paths emanating from an atom to an atom-complexity from which a number of molecular complexity measures are derived. The method is independent of explicitly predefined features such as ring membership, bond types, chirality or symmetry. These path-based measures of complexity can distinguish subtle differences in molecular structure and an application to the visualization of marketed drugs, including a number of biologics, is presented.

There are many references to molecular complexity in the chemistry literature and several methods are available to measure it. Graph and information theoretical methods were developed by Randic,[1] Bonchev,[2] Bertz,[3] Hendrickson,[4] Rucker[5] Kier[6] and von Korff.[7] Methods based on the occurrence of selected molecular features or substructures were provided by Boda and Johnson,[8] Whitlock,[9] Barone and Chanon,[10] Oprea[11], and Ertl.[12] More recently, Bottcher provided an additive atom-based approach.[13] Collective intelligence[14] and crowdsourcing[15] approaches have also been reported. Some of these methods have been used to analyze synthesis routes and Sarpong recently used a measure of complexity to prospectively identify key bridgehead disconnection points in a synthesis of weisaconitine.[16]

As noted by Shannon, "quantities of the form $H = -\sum p_i \log p_i$ where p_i represents the fractional occurrence of an invariant N in a system play a central role in information theory as measures of information, choice and uncertainty".[17] Applications of this equation provide measures of graph complexity and, by extension, the complexity of the 2-D graphical representation of molecules. In Shannon's formulation p_i represented the fractional occurrence of symbols as invariant classes within a message. Bond and atom types are corresponding invariants for molecular complexity.

Beyond the application of Shannon's concept to calculating complexity in a whole-molecule sense, it is also possible to define a complexity for each individual atom environment. In this case, the number and variety of paths emanating from an atom determine the complexity of that atom environment. The application of Bertz's formula[3] to atoms rather than molecules, as in Figure 1 (a), provides the complexity C_A, of each atom environment. Here p_i is the fractional occurrence of each path type emanating from a particular atom and N is the total number of paths emanating from that atom.

(a) $C_A = -\sum p_i \log_2 p_i + \log_2 N$

(b) $C_M = \sum C_A$

(c) $C_{M'} = \log_2 \sum 2^{C_A}$

(d) $C_{SE} = -\sum q_i \log_2 q_i$

Figure 1. The basic equations for complexity calculations

Molecular complexity can be defined as either the simple sum of the C_A, C_M, as in Figure 1 (b) or the log-sum of the exponentials[18] of the C_A, C_M*, Figure 1 (c). This latter summing has the property of emphasizing the contribution of the most complex feature(s) to the total complexity.

We also derive C_{SE} from the fractional occurrence of atom types, q_i in equation (d), and this gives a measure of the diversity of atom environments in the molecule.

The details of molecular representation, path length and atom properties as set for the calculation of the atom paths are as follows. We follow the convention of using H-suppressed molecular graphs and ignore paths emanating from H-atoms. We