Classroom

In this section of *Resonance*, we invite readers to pose questions likely to be raised in a classroom situation. We may suggest strategies for dealing with them, or invite responses, or both. "Classroom" is equally a forum for raising broader issues and sharing personal experiences and viewpoints on matters related to teaching and learning science.

Investigation of Structures Similarity of Organic Substances

Ajay Kumar
Guru Tegh Bahadur Institute of Technology
G–8 Area, Rajouri Garden
New Delhi 110064, India.
Email:
ak.gupta59@rediffmail.com

According to similarity property principle, structurally similar molecules tend to have similar properties. Similar molecules exhibit similar biological activities. However, there is no hard and fast rule that the compounds with similar chemical structure will have similar functions. There are several compounds of similar chemical structure with significantly different biological actions and activities. Structure similarity of natural products can be investigated using Tanimoto coefficient and Euclidean distance measurements. To do this, the molecules are decomposed into smaller fragments, and a dictionary, of the fragments are prepared. Using the dictionary, 2D fingerprints are formed. With the fingerprints Tanimoto coefficient and Euclidean distance are quantitated.

Keywords
Structures similarity, Tanimoto coefficient, Euclidean distance, fingerprints (bit-string representations).

1. Introduction

The study of the relationship between molecular structure and molecular function or reactivity of the substances (particularly of the organic compounds) is integral to chemistry. According to 'similarity property principle' [1], "the structurally similar molecules tend to have similar properties and similar molecules exert similar biological activities". Medicinal chemists have made use of this concept to modify the structures of biologically active

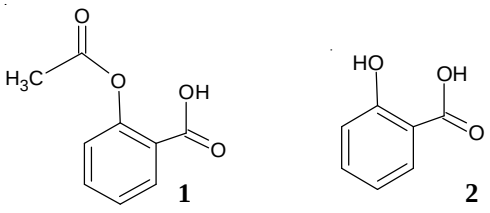
natural products using Tanimoto coefficient and Euclidean distance methods [6–8]. Smaller molecules may appear to be structurally closer when using the Euclidean distance, even if the absence of common features are taken into account while measuring Euclidean distance [7].

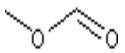
The structures of the two substances in *Table* 1 look similar. Each contains an aromatic ring and a > CO group. Substance **1** is an ester and substance **2** contains an -OH group. Both substances **1** and **2** contain a carboxylic group.

The Method

Tanimoto Coefficient: The structures of the substances to be compared are decomposed into fragments and the structural fragments or features that are present in the given molecule are turned ON (set as 1) and the ones that are absent are kept OFF (set as 0), [9]. Thus, for each molecule there will be a string containing 1s and 0s (bit string) as determined by the elements of the dictionary. Bits are set only once irrespective of the frequency of occurrence of the given key. Bits get set on the basis of fragments of whole structure and there is poor capturing of properties of the whole molecule. The structures [10] of the two natural products **1** and **2** under consideration are decomposed into smaller fragments and a dictionary of all the fragments is prepared (*Table* 1). The

Table 1. Structures and fragments of two molecules and their string containing 1s and 0s (bit string) of the natural products.



1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0
Fragments								

fragments present in the dictionary are matched with the fragments of the individual natural product. If a fragment of the natural product matches with a fragment present in the dictionary, it is given the value of '1' and the fragment that is absent is given the value of '0'. This way a bit-string fingerprint of each natural product is prepared (*Table* 1).

Once the bit-string fingerprint is ready, any of the association coefficients can be determined to find out the similarity between any two given molecules. Tanimoto coefficient is widely used to determine the similarity based on fingerprints (bit-string representations). For example, the Tanimoto coefficient (τ) for two molecules **1** and **2** (*Table* 1) is determined by comparing bit-string fingerprints using the equation:

$$\tau = N_{12} / N_1 + N_2 - N_{12} ,$$

where N_1 is the number of features (ON bits) in **1**, N_2 is the number of features (ON bits) in **2**, and N_{12} is the number of features (ON bits) common to both **1** and **2.**

$$\tau = 3/ (5 + 3 - 3) = 0.6 .$$

Tanimoto coefficient is also known as Jaccard Coefficient. 1s are treated as significant here and 0s are treated as not significant. Similarity varies between 0 (dissimilar) and 1 (same). A good cut-off for biologically similar molecules is 0.7 or 0.8

Euclidean Distance Measurement: Euclidean distance measurement is Pythagorean distance and has binary dimensions equivalent to the square root of the Hamming distance (i.e., square root of the number of bits that are different). 0s are treated as significant. Smaller values mean more similarity [9].

Example:

	↓ ↓
	1101**1**0**1**0
	1101**0**0**0**0
Different	x x

Euclidean distance = $\sqrt{2}$.

Results

The values of the similarity coefficients show that substance **1** and substance **2** have similarity with Tanimoto coefficient of 0.60 and Euclidean distance of $\sqrt{2}$. However, a good cut-off of Tanimoto coefficient for likely biological similarity of the molecules is 0.7 or 0.8. Higher the value of Tanimoto coefficient higher is the similarity. Lower the value of Euclidean distance higher is the similarity. Therefore, since the above two compounds have low similarity, they may show similar biological activities.

Conclusions

Using the above similarity coefficient methods both the undergraduate and the post-graduate students can investigate the similarity or dissimilarity of natural products or other substances.

Suggested Reading

[1] **A M Johnson and G M Maggiora, *Concepts and Applications of Molecular Similarity*, Ist Ed, New York, John Wiley & Sons, 1990.**

[2] **C Hansch, P G Sammes and J B Taylor, *Comprehensive Medicinal Chemistry*, Pergamon Press, Oxford, 1990.**

[3] **M E Wolff, *Burger's Medicinal Chemistry*, 5th ed., John Wiley & Sons, New York, 1995.**

[4] **P M Dean, *Molecular Similarity in Drug Design*, Blackie Academic & Professional, London, 1995.**

[5] **C G Wermuth, *The Practice of Medicinal Chemistry*, Academic Press, London, 1996.**

[6] **D R Flower, On the Properties of Bit String-Based Measures of Chemical Similarity, *J. Chem. Inf. Comput. Sci.*, Vol.38, No.3, pp.379–386, 1998.**

[7] **P Willett, J M Barnard and G M Downs, Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.*, Vol.38, No.6, pp.983–996, 1998.**

[8] **K Oliver, *2D Similarity*, www-bs.informatik.unituebingen.**

[9] **A R Leach and V J Gileet, *Introduction to Chemo informatics*, Ist Ed, Kulwer Academic Publishers, USA, 2003.**

[10] **M Windholz and N J Rathway, *The Merck Index*, 10th Ed., Merck & Co. Inc., 1983.**