28-29 September 2023,
Kragujevac, Serbia

www.iccbikg2023.kg.ac.rs

2nd International Conference on Chemo and Bioinformatics
ICCBIKG_2023

ICCBIKG
International Conference on Chemo and BioInformatics

BOOK OF PROCEEDINGS

2nd International Conference on Chemo and BioInformatics

ICCBIKG 2023

BOOK OF PROCEEDINGS

September 28-29, 2023

Kragujevac, Serbia

Sponsored by

2nd International Conference on Chemo and BioInformatics, Kragujevac, September 28-29, 2023, Serbia.

Editors:

Professor Dr. Zoran Marković

Professor Dr. Nenad Filipović

Technical Editors:

Jelena Živković

Marko Antonijević

Dr. Žiko Milanović

Dr. Vladimir Simić

Proofreading:

Marijana Dimović

Publisher:

Institute for Information Technologies, University of Kragujevac, Serbia, Jovana Cvijića bb, 2023

Press:

„Grafo Ink“, Kragujevac

Impression:

120 copies

CIP - Каталогизација у публикацији - Народна библиотека Србије, Београд

54:004(048)(0.034.2)
57+61]:004(082)(0.034.2)

INTERNATIONAL Conference on Chemo and BioInformatics (2 ; 2023 ;
Kragujevac) Book of Proceedings [Elektronski izvor] / 2nd International Conference on
Chemo and BioInformatics, ICCBIKG 2023, September 28-29, 2023 Kragujevac,
Serbia ; [editors Zoran Marković, Nenad Filipović]. - Kragujevac : University,
Institute for Information Technologies, 2023 (Kragujevac : Grafo Ink). - 1 USB
fleš memorija ; 1 x 2 x 6 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovne strane dokumenta. - Tiraž
120. - Bibliografija uz svaki rad.

ISBN 978-86-82172-02-4

a) Хемија -- Информациона технологија -- Зборници b) Биомедицина --
Информациона технологија -- Зборници

COBISS.SR-ID 125908489

A metric for pairwise similarity analysis of binary cheminformatics data

Izudin Redžepović[1,*]

[1] State University of Novi Pazar, Vuka Karadžića 9, 36300 Novi Pazar, Serbia; e-mail: iredzepovic@np.ac.rs

* *Corresponding author*

DOI: 10.46793/ICCBI23.593R

Abstract: This paper unveils the findings derived from an in-depth exploration of a novel similarity measure designed to assess pairwise resemblances. Called the Substructure Similarity Index, this measure centers around the comparison of substructures identified within compounds. Through a rigorous evaluation conducted on an extensive dataset of drugs and by juxtaposing it against other commonly employed indices, the study reveals that the Substructure Similarity Index can be adeptly employed for molecular similarity calculations since it provides information that cannot be obtained by available measures.

Keywords: molecular similarity, molecular structure, binary vectors, molecular fingerprints, similarity measure

1. Introduction

Countless chemical procedures have been developed to create molecules that resemble existing ones but exhibit specific, highlighted characteristics. However, the significance of molecular similarity extends beyond the realm of chemistry and finds implementation in diverse fields, including pharmaceuticals, materials design, agriculture, toxicology, and many more [1]. For example, computer-assisted retrosynthesis uses molecular similarity to assist in planning synthetic routes [2], while mass spectrometry-based annotation of natural product compound families relies on it to identify related compounds [3]. These uses underscore the wide-ranging scope and potential impact of molecular similarity across different disciplines.

At the heart of molecular similarity lies the principle that structural similarity leads to similar activity [4]. The significance of choosing different similarity metrics and their impact on the variability of similarity assessment results have been widely acknowledged [5]. In this work, we introduce a novel pairwise similarity measure for binary vectors in order to eradicate the shortcomings of existing measures. To evaluate its performance, we conducted a comparative statistical analysis.

2. Theory

In this part, we define Substructure Similarity Index (*SSI*) and discuss its basic properties. The *SSI* is defined as follows:

$$SSI = \frac{ax + by}{n + m} \tag{1}$$

whereas a and b are the length of the longest string of ones in the fingerprint A (i.e., the substructure of molecule A) and the length of the longest string of ones in the fingerprint B (i.e., the substructure of molecule B), respectively. The x and y denote how many times a and b occur, respectively. The n and m stand for the number of bits one in binary vectors A and B, respectively. From Eq. (1), it is obvious that *SSI* yields similar results in the [0,1]-range. Note that the *SSI* allows the comparison of vectors with different lengths. It is worth emphasizing that *SSI* compares only the biggest substructures of two molecules, preventing false similarity caused by small structural overlaps.

2.1 Molecular library and computational details

To investigate *SSI*, we have performed a comparative statistical analysis of our metric with other similarity indices. More precisely, for this purpose, Tanimoto (*T*), Jaccard (*Ja*), Gleason (*Gle*), Sokal-Sneath (*SS*), and Consonni-Todeschini (*CT*) indices have been employed. In this work, as a case study, we have used the FDA approved drugs available in DrugBank database V. 5.1.10, that consists out of 975 compounds. The Morgan circular fingerprint (1024 bits, radius=2) has been used to represent the chemical structure of each compound.

3. Results and Discussion

The statistical parameters corresponding to the similarity values computed using six distinct similarity measures are outlined in Table 1. Moreover, the visual representation of their distributions can be observed in Figure 1. The first five indices all exhibit a minimum value of zero, in contrast to the *SSI*, which demonstrates a value of 2.38%. This distinctive behavior of *SSI* underscores its superiority among the other measures. Specifically, this dataset comprises a variety of molecules that possess certain structural elements in common, including identical atoms and bonds, as a minimum. So, having zero similarity between some molecules is questionable. A visual representation of this concept is depicted in Figure 2, providing a clear illustrative example. Among all the indices, *CT* consistently yields the highest average values, while *SSI* displays the most pronounced data scattering. This observation regarding *SSI* aligns with the notable diversity present within the compound set.

Table 1. Statistical parameters for the similarity indices: minimal (min), maximal (max), and the mean value (all in %), and the *s* denotes standard deviation.

Measure	min	max	mean	*s*
T	0	100	10.37	5.17
Ja	0	100	24.95	9.75

Gle	0	100	18.43	7.94
SS	0	100	5.55	3.17
CT	0	100	47.65	11.59
SSI	2.38	100	32.62	28.63

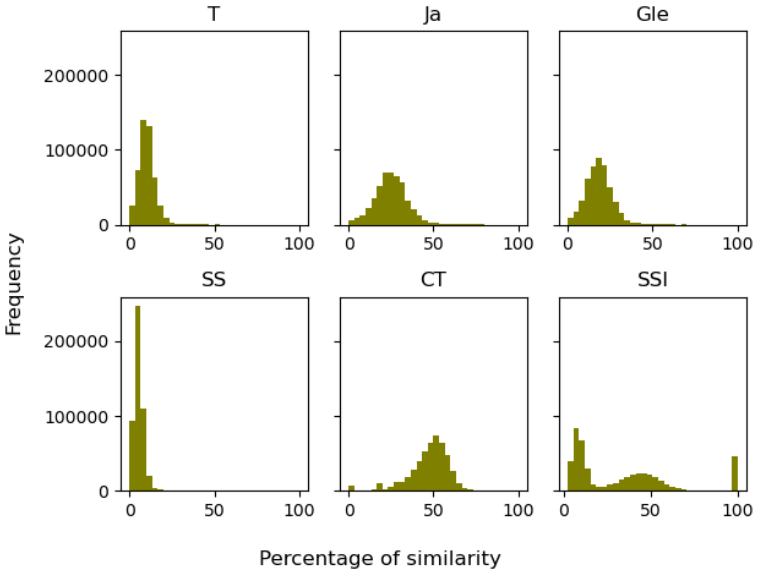

Figure 1. Distribution of 474 825 similarity values calculated by six different similarity measures.

To explore the correlation between *SSI* and other indices, we have computed the Pearson correlation coefficients between the similarity measures. The outcomes of these calculations are showcased in Table 2. As one may see, it is evident that *SSI* shows negligible correlation with other indices. This observation substantiates the rationale behind introducing *SSI*, as it has the potential to offer insights into molecular similarity that existing measures cannot capture.

Table 2. The absolute values of the correlation coefficient between similarity indices.

	T	*Ja*	*Gle*	*SS*	*CT*	*SSI*
T	1					
Ja	0.9762	1				
Gle	0.9901	0.9968	1			
SS	0.9880	0.9344	0.9579	1		
CT	0.8643	0.9305	0.9088	0.8082	1	
SSI	0.0455	0.0604	0.0540	0.0397	0.1927	1

The comprehensive findings presented in this study collectively affirm the efficacy of employing *SSI* for molecular similarity calculations. Notably, its definition suggests that *SSI* holds promising potential for successful utilization in substructure searching applications.

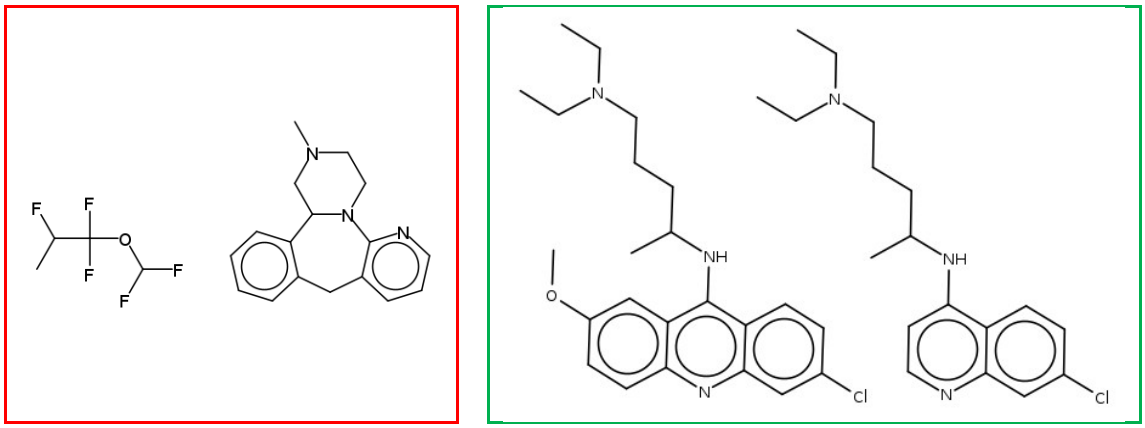

Figure 2. Red box: Two least similar molecules within the dataset. The *SSI* shows 2.38% similarity between two compounds. Other indices estimate it at 0, even though they share some structural details, like the C-C bonds. Green box: A pair of molecules for which *SSI* shows 100% similarity, while other indices yield significantly less percentage of similarity.

4. Conclusions

The Substructure Similarity Index introduces a fresh approach to quantifying pairwise similarity within binary vectors. The outcomes underscore its potential significance in similarity assessments, as it imparts novel insights not covered by existing measures.

Acknowledgment

Izudin Redžepović thanks the Serbian Ministry of Science, Technological Development, and Innovation for its support (Grant No. 451-03-47/2023-01/200122). The author also acknowledges financial support from the State University of Novi Pazar.

References

[1] A. Bender, R.C. Glen., *Molecular Similarity: A Key Technique in Molecular Informatics*, Organic & Biomolecular Chemistry, 2 (2004) 3204-3218.

[2] C.W. Coley, L. Rogers, W.H. Green, K.F. Jensen., *Computer-Assisted Retrosynthesis Based on Molecular Similarity*, ACS Central Science, 3 (2017) 1237-1245.

[3] N.J. Morehouse, T.N. Clark, E.J. McMann, J.A. van Santen, F.P.J. Haeckl, C.A. Gray, R.G. Linington., *Annotation of Natural Product Compound Families Using Molecular Networking Topology and Structural Similarity Fingerprinting*, Nature Communications, 14 (2023) #308.

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

[5] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, *Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets*, Journal of Chemical Information and Modeling, 52 (2012) 2884-2901.