Similarity Measure for Molecular Structure: A Brief Review

S A Bero, A K Muda, Y H Choo, N A Muda and S F Pratama
Computational Intelligence and Technologies (CIT) Research Group, Center of Advanced Computing and Technologies, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

sitiasmah.bero@gmail.com

Abstract. Similarity or distance measures have been used widely to calculate the similarity or dissimilarity between two samples of dataset. Cheminformatics is known as the domain that dealing with chemical information and both similarity and distance coefficient have been an important role for matching, searching and classification of chemical information. There are various types of similarity/distance coefficient used in molecular structure similarity searching. Generally, the calculation using similarity/distance coefficient is focusing more on 2-dimensional (2D) rather than 3-dimensional (3D) structure. In this paper, the popular similarity/distance coefficients for molecular structure will be reviewed together with the review on 3D molecular structure.

1. Introduction
The Cambridge English Dictionary defines similar to be “looking or being almost, but not exactly, the same” and the comparison of objects to determine their levels of similarity lies at the heart of many academic disciplines. The “Similarity” also said as the resemblance, likeness, sameness and an agreement. The concept of similarity searching was being applied to numerous types of application and domain, such as in pattern recognition, chemistry and cheminformatics. The high demand for drug discovery research recently makes the molecular structure searching is one of the prominent research areas in the cheminformatics.

The earliest study of similarity searching on molecular structure started in mid-eighties that was conducted by [1] Raymond et al., that proposed an algorithm for computing the simple type of substructure similarity known as atom pairs. A year later, the similarity studies are continued by [2, 3] until now.

The similarity searching of molecular structure is closely related to the Similar Property Principle, whereby, molecules that are structurally similar are likely to have similar properties [4]. This relationship lead to the range of cheminformatics techniques as it had established a discipline that has applications in many areas of chemistry and computational molecular science [5, 6] such as molecular diversity analysis, clustering, QSAR/QSPR methods and also similarity searching [4].

In order to conduct the similarity searching, there are three important components of similarity measures that need to be considered [7]. First, the structure representation (used to characterize molecules), second, the weighting scheme (used to determine the relative importance value of the representation), and lastly is similarity coefficient. The similarity coefficient plays an important role...
in the context of similarity search, as it is used to quantify the degree of resemblance between the two structural representations [4].

The molecular representation is always being the first thing to consider before conducting any similarity searching in molecular structure. There are various types of descriptors reported [8-10] that had been used in the studies of molecular similarity. According to [4], the descriptors can be divided into three classes: 1D, 2D and 3D representations.

A molecular structure consists of a chemical formula that indicates the numbers and types of atoms in a molecule. Whereby, each atom is linked together by a bond to form a complete molecule structure. A molecular structure is an important element in the area of similarity searching as it is bringing the description of type, arrangement, position and direction of the bonds linking atom within a molecule [11]. Besides that, the molecular structure is the fundamental unit in studying the chemical information of molecule. Whereby, variety types of chemical compounds will uniquely identify by their molecular structure [12].

A molecular structure can be expressed in either 1D, 2D or 3D [13]. Many researchers have worked with 2D molecular structure rather than 3D structure, for example, in the study of molecular structure similarity searching [14-17]. This is due to the fact that 2D structure is simpler, easy to work on, and not as complicated as 3D. A molecular structure becoming a main component in the drug discovery whereby that chemical information extracted from the structure is very valuable for the chemist.

Furthermore, a molecular structure been an important aspect in drug discovery, because many kind of drugs can be identified based on their consistent structure of its molecular structure. Although there are many research has been done on the 2D molecular structure, however less attention is given to 3D structure similarity searching. Therefore, the study of 3D structure similarity searching is still relevant and open for exploration as the 3D structure will capture all aspects of molecular structure that 2D does not, which are the size of molecular structure and the shape of molecular structure.

This paper presents on some of the popular similarity/distance measurer used for similarity searching of molecular structure. At the same time, this paper also reviews on 3D molecular structure. The paper will be organized as follows: Section 2 introduces the 3D molecular structure, followed by an analysis and a timeline of the similarity measurer in Section 3. Finally, the conclusions are given in Section 4.

2. 3D Molecule Structure

A chemical structure can be represented either in 2D or 3D, as shown in figures 1 and 2 respectively. Before the emergence of 3D structure, the 1D and 2D has been widely used in previous research on chemical structure; especially in similarity search of chemical structure. Those structures become a priority previously because both 1D and 2D structure allows for ease of developing an algorithm [18] and easy to work on. Later, this section will briefly discuss on the characteristic of 3D molecular structures in the context of similarity searching.

![Figure 1](image.png)

**Figure 1.** An example of 2D molecule structure of amphetamine
2.1. Characteristics of 3D Chemical structure

According to [19], the molecular similarity is a complex structure that can only be described with reference to the immediate use for which it is intended. Similarity encompasses the bonding patterns, atomic positions, conformation, shape and spatial disposition of molecular properties [19]. This is simply saying that, to undergo a 3D similarity chemical structure, it is needed to consider the whole properties of 3D chemical structure itself.

In the context of 3D similarity searching, it requires the conformation properties and the specification of an entire target structure (the whole structure) rather than partial structure. Thus, the 3D method-based similarity measurement was introduced and gained more attention recently because of their potential to overcome the key limitation of 1D and 2D methods [13].

The 3D structures usually have three different angles of view, which is $x$, $y$, and $z$-axis. Meanwhile, the 2D structure only has two angles of view; $x$- and $y$-axis. With the one extra angle of view in 3D structure, it will provide an important information needed and better conformation for the similarity search process. Another characteristic of 3D structure is, it is better in capturing all aspects of the molecular structure (the size of molecular structure and the shape of molecular structure). In order to measure how similar is the chemical structure is, a similarity measure is needed.

3. Similarity and Distance

A similarity or distance measure can be defined as a tool to quantify the similarity or dissimilarity between object or known as a function that can compute the degree of similarity or dissimilarity between a pair of object. Numerous similarity measures and distance measures have been used widely in various fields of the area. For example, in social network [20], text similarity [21], document similarity [22], triangle inequality [23], biology [24, 25], image retrieval [26-28], fingerprint [29, 30], handwritten character [31, 32], and chemistry [14-17, 33].

From mathematical point of view, distance is used to define how far between two objects are [34]. Meanwhile, from the metric point of view, the distance also known as dissimilarity, whereby the concept is still the same; to find the distance between two objects. In terms of distance coefficients, the distance is used to quantify the degree of difference between two objects. The smaller value of distance, the larger the degree of similarity and vice versa [35].

Figure 3 below shows the similarity measure and distance measure that is commonly used in 2D binary feature vector that was arranged in historical order [36]. Those measures are divided into three different classes; one class of distance based, and two classes of similarity measure (non-correlation based and correlation based).

Based on figure 3, the majority of the similarity measure is grouped under the Correlation based. Jaccard, Russell & Rao, Dice & Sorenson, Tanimoto, Roger, Sokal & Michener, Hamann, Faith and Gower & Legendre are grouped under Non- Correlation coefficient in [36]. Meanwhile, this Non-Correlation coefficient is grouped as an Association coefficient in [35] and there are 16 of them,
including Jaccard/Tanimoto, Dice, Russell/Rao, Sokal/Sneath (1)-(3), Kulczynski (1)-(2), Simple matching, Hamann, Rogers/Tanimoto, Baroni-Urbani/Buser, Ochiai/Cosine, Forbes, Fossum and Simpson.

3.1. Past Studies of 3D Similarity Searching
The 3D molecular structure similarity search is believed to overcome the limitation of 1D and 2D structure on the aspect of retrieving important information which cannot be retrieved on both 1D and 2D structures. Hence, the 3D molecular structure similarity search becomes a concern nowadays, especially in the drug discovery.

![Figure 3. The chronological of binary similarity and distance measures by year [36]](image)

The 3D molecular structure similarity searching has begun since more than 2 decades ago and a numerous method and approach has been introduced. For example, in the early 1980s, [37] had done a matching measure between two molecular structures based on the density function, and followed by [38], which used an application namely SPERM (Superposition by PERMutations), to search a 3D molecular structure of two biological molecules namely netropsin and daunomycin in 30 000 compound databases.

Meanwhile, [39] have proposed four different algorithms of pattern matching to determine the presence or absence of an atom and associated interatomic distance in 3D chemical structure and the results show that, out of four algorithms proposed, the Ullman algorithm is more efficient on small molecular structure searching compared to the others.

The study of similarity searching for 3D molecular structure represented by interatomic distance matrices is continued in the year of 1990s. The study done by [40] had compared between atom mapping and clique detection method for the calculation of 3D structural similarity. From the experiment, the atom mapping method is more efficient in small 3D molecular similarity searching, compared to the Clique detection is well measured for 3D macromolecular similarity searching. Later, the study in 3D molecular structure similarity searching keeps rising year by year as reported by [14, 41-44].
Nowadays, the similarity searching of 3D molecular structure had rapidly developed with various applications and methods which are reviewed in [13]. Commonly, the 3D representation was measured based on their shape similarity [45], which compared to 2D similarity, most similarity calculation is based on 2D fingerprints (0 and 1) [4]. The similarity/distance coefficient commonly used in 2D, also being used in 3D. For example, Manhattan distance, Tanimoto coefficient, and cosine coefficient [13].

### 3.2. Analysis of Similarity Measurement for Molecular Structure

According to [7], a similarity searching will be effective if, and only if, an appropriate similarity measure is employed. Of the similarity measure listed in the previous study, a similarity measure under an association coefficient is commonly used in similarity searching of molecular structure, for example, the Tanimoto coefficient [14]. The association coefficient is specifically designed for use with binary data, which used 0 and 1 to indicate the absence and the presence of molecular structure [33].

| Table 1. Common similarity/distance measure in chemical informatics |
|---------------------------------------------------------------|
| Similarity/Distance Coefficient | Formula |
| Tanimoto Coefficient | \( Tani = \frac{c}{a+b-c} \) (1) |
| Dice Coefficient | \( Dice = \frac{2a}{2a+b+c} \) (2) |
| Cosine Coefficient | \( Cos = \frac{a}{\sqrt{(a+b)(a+c)}} \) (3) |
| Euclidean Distance | \( ED(a,b) = \sqrt{\sum_{i=1}^{n} (b_i-a_i)^2} \) (4) |
| Hamming Distance | \( HM(a,b) = a + b - 2c \) (5) |

The nature of association coefficient usually working with a binary data and often normalized between the range of zero (no similarity at all) and unity (identical set of descriptors) [33, 35]. Besides the binary data, the association coefficient also can be used with non-binary data, in which the different formulation and range of value may apply [35]. Tanimoto coefficient, Dice coefficient, Cosine coefficient, Euclidean distance and Hamming distance also being known as the most similarity measures used in similarity searching of molecular structure studies.

Equation (1) refers to the Tanimoto coefficient (also known as the Jaccard coefficient) and it said as the extension from the Jaccard coefficient [16]. The Tanimoto coefficient is defined as the sum of the common feature occur in both structures (c) over the total feature that occur only in a and b, minus with the common feature, c.

Most of the previous research by research group at the University of Sheffield used the Tanimoto coefficient as the similarity measure in lots of their experiment on similarity searching of the chemical structure as presented in [2-4, 7, 9, 14, 33]. The reason why the Tanimoto coefficient becomes a priority in the context of similarity of molecular structure was presented in [40]. The Tanimoto coefficient was preferred based on two reasons; firstly, on the basis of a subjective evaluation of the similarity search ranking, and secondly, referred to its calculation that does not involve any square root, making it faster in calculation [14].

The ease of implementation and speed is the main reason why a Tanimoto coefficient becomes the most popular similarity measure in the context of chemical informatics and computational medicinal chemistry [46]. Besides that, the other similarity measure such as Dice coefficient, Cosine coefficient,
Euclidean distance and Hamming distance also being used in similarity searching of molecular structure but not rapidly as the Tanimoto coefficient. The Dice coefficient in equation (2) also referred as the Sorensen, Czekanowskli, Hodgkin-Richard or Morisita [34]. The Dice coefficient measure the similarity directly like Tanimoto coefficient [10] making it ease in implementation and fast in the calculation.

Equation (3) indicates the Cosine coefficient that is used to measure the similarity between two vectors by measuring the cosine of the angle between the data objects and then, the similarity is measured as the angle between them [47]. Compared to Tanimoto and Dice coefficient, the Cosine coefficient is often used to measure the similarity between sparse data (like document or binary data). For example, text mining and measure the cohesion within clusters in the field of data mining [47]. In the context of chemical similarity, the Cosine coefficient allows the calculation of the average similarity between all pairs of compound in the dataset very rapidly which is impossible when using Tanimoto coefficient [14].

Meanwhile, the Euclidean distance and Hamming distance are distance coefficient that been used to quantify the degree of dissimilarity between two molecules. Both of these coefficients effectively in identifying the common absence of the feature as evidence of similarity [10], and [14] stated that, the Hamming distance is equivalent to the squared Euclidean distance for binary data. Meanwhile, the Euclidean distance is very useful for measuring the similarity between dense data such as time series or two dimensional points [47]. It is reported as the Euclidean distance remains the most popular measure of distances either for continuous data or count data [16].

4. Conclusion
This paper reviewed the common similarity/distance measurer in molecular structure. The Tanimoto coefficient is found the most widely similarity measures used in similarity searching of chemical informatics fields compared to the other similarity measure. Besides that, the majority of the similarity and distance measurer is commonly used for binary calculation that is referred more on 2D chemical structure. Meanwhile, the similarity searching of 3D chemical structure must dealing with their conformational properties and shape representation in order to find the similarity. Some researchers believed that the similarity and distance measurer can be used for non-binary data (especially on the association coefficient), but the modification to the range of the value is need. In conclusion, the studies of similarity searching in chemical informatics offer the advantages to the chemist and also in the field of drug discovery itself.

Acknowledgments
This work is funded by Universiti Teknikal Malaysia Melaka (UTeM) through the PJP High Impact Research Grant [PJP/2016/FTMK/HI3/S01473].

References
[1] Carhart R E, Smith D H, and Venkataraghavan R 1985 Atom pairs as molecular features in structure-activity studies: definition and applications (J. Chem. Inf. Comput. Sci. vol 25) chapter 2 pp 64–73
[2] Willett P, Winterman V, and Bawden D 1986 Implementation of nearest-neighbor searching in an online chemical structure search system (J. Chem. Inf. Model vol 26) pp 36–41
[3] Willett P and Wintermann V 1986 A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity Measures of Inter-Molecular Structural Similarity (Quant. Struct. Relationships vol 25) pp 18–25
[4] Willett P 2014 The calculation of molecular structural similarity: Principles and practice (Mol.Inform. vol 33) chapter 6–7 pp 403–413
[5] Arulmozhi, Chemoinformatics V 1907 A Quick Review pp 416–419
[6] Begam B F and Kumar J S 2012 A Study on Cheminformatics and its Applications on Modern Drug Discovery (Procedia Engineering 38(Icmoc)) pp 1264–1275
[7] Willett P 2009 Similarity methods in chemoinformatics (Annu. Rev. Inf. Sci. Technol. vol 43), pp 1–117
[8] Todeschini R and Consonni V 2002 Handbook of Molecular Descriptors Wiley-VCH, Weinheim
[9] Warr W E 2011 WIREs Computational Molecular Sciences vol 1 pp 557
[10] Leach A R and Gillet V J 2007 An introduction to Chemoinformatics
[11] Molecular structure. (n.d.) Jonas: Mosby's Dictionary of Complementary and Alternative Medicine. (2005), http://medicaldictionary.thefreedictionary.com/molecular-structure
[12] Chemical structure, https://en.wiktionary.org/wiki/chemical_structure
[13] Shin W H, Zhu X, Bures M, and Kihara D 2015 Three-Dimensional Compound Comparison Methods and Their Application in Drug Discovery (Molecules vol 20) chapter 7 pp 12841–12862
[14] Willett P, Barnard J M and Downs G M 1998 Chemical similarity searching (Journal of chemical information and computer sciences vol 38) chapter 6 pp 983–996
[15] Hagadone T R 1992 Molecular substructure similarity searching: efficient retrieval in two dimensional structure databases (Journal of chemical information and computer sciences vol 32) chapter 5 pp 515-521
[16] Willett P 2006 Similarity-Based Virtual Screening Using 2D Fingerprints (Discovery vol 1) (New York: Universities of Leeds, Sheffield and York) pp 1046–1053
[17] Flower D R 1998 On the Properties of Bit String-Based Measures of Chemical Similarity (Journal of Chemical Information and Modeling vol 38) chapter 3 pp 379–386
[18] Shin W H, Zhu X, Bures M G and Kihara D 2015 Three-dimensional compound comparison methods and their application in drug discovery (Molecules vol 20) chapter 7 pp 12841-12862
[19] Dean P M 2012 Molecular Similarity in Drug Design vol 65 chapter 3
[20] Rawashdeh A and Ralescu A L 2015 Similarity Measure for Social Networks-A Brief Survey In MAICS pp 153-159
[21] Mihalcea R, Corley C and Strapparava C 2006 Corpus-based and knowledge-based measures of text semantic similarity (In AAAI vol 6) pp 775-780
[22] Huang A 2008 Similarity measures for text document clustering (In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand) pp 49-56
[23] Lipkus A H 1999 A proof of the triangle inequality for the Tanimoto distance (Journal of Mathematical Chemistry vol 26) chapter 1-3 pp 263-265
[24] Hubalek Z 1982 Coefficients of Association and Similarity, Based on Binary (Presence Absence) Data: An Evaluation (Biological Reviews vol 57) chapter 4 pp 669–689
[25] Gray K R, Aljabar P, Heckemann R A, Hammers A and Rueckert D 2013 Alzheimer's Disease Neuroimaging Initiative.: Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. (NeuroImage vol 65) pp 167-175
[26] Smith J R and Chang S F 1996 Automated binary texture feature sets for image retrieval (International Conf. Accoust., Speech, Signal processing, Atlantic, GA)
[27] Subrahmanya M, Maheshwari R P and Balasubramanian R 2012 Local maximum edge binary patterns: a new descriptor for image retrieval and object tracking (Signal Processing vol 92) chapter 6 pp 1467-1479
[28] Cerra D and Datcu M 2012 A fast compression-based similarity measure with applications to content-based image retrieval (Journal of Visual Communication and Image Representation vol 23) chapter 2 pp 293-302
[29] Ghany K K A, Hassanien A E and Schaefer G 2014 Similarity measures for fingerprint matching (In International Conference on Image Processing, Computer Vision, and Pattern Recognition, USA) pp 21-24
[30] Bajusz D, Rácz A and Héberger K. 2015 Why is Tanimoto index an appropriate choice for
fingerprint-based similarity calculations? (Journal of cheminformatics vol 7) chapter 1 pp 1
[31] Rodriguez-Serrano J A and Perronnin F 2012 A model-based sequence similarity with application to handwritten word spotting (IEEE transactions on pattern analysis and machine intelligence vol 34) chapter 11 pp 2108-2120
[32] Wang Q F, Yin F and Liu C L 2012 Handwritten Chinese text recognition by integrating multiple contexts (IEEE transactions on pattern analysis and machine intelligence vol 34) chapter 8 pp 1469-1481
[33] Willett P 2011 Similarity searching using 2D structural fingerprints (Chemoinformatics and computational chemical biology) pp 133-158
[34] Cha S 2007 Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions (Int. J. Math. Model. Methods Appl. Sci vol 1) chapter 4 pp 300–307
[35] Screening H T 2002 Grouping Of Coefficients For The Calculation Of Inter-Molecular Similarity And Dissimilarity Using 2D Fragment Bit-strings vol 5(New York: Universities of Leeds, Sheffield and York) pp 155–166
[36] Seung-Seok C, Sung-Hyuk C and Tappert C 2010 A Survey of Binary Similarity and Distance Measures (Journal of Systemics, Cybernetics & Informatics vol. 8) chapter 1 pp 43–48
[37] Carbó R, Leyda L and Arnau M 1980 How similar is a molecule to another? An electron density measure of similarity between two molecular structures (Int. J. Quantum Chem., vol 17) chapter 6 pp 1185–1189
[38] Van Geerestein V J, Perry N C, Grootenhuis P D J and Haasnoot C A G 1990 3D Database searching on the basis of ligand shape using the sperm prototype method (Tetrahedron Computer Methodology vol 3) chapter 6 PART C pp 595–613
[39] T Brint A and Willett P 1987 Pharmacophoric pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms (Journal of Molecular Graphics vol 5) chapter 1 pp 49–56
[40] Artymiuk P J, Bath P A, Grindley H M, Pepperrell C A, Poirrette A R, Rice D W, Thorner D. A, Wild D J, Willett P, Bank W, Sheffield S, Allen F H, and Taylor R 1992 Similarity Searching pp 617–630
[41] Bath P A, Poirrette A R, Willett P, Sheffield S and Allen F H 1994 Similarity Searching in Files pp 141–147
[42] Rouvray D H 1994 Similarity studies. 1. The necessity of analogies in the development of science( Journal of Chemical Information and Modeling vol 34) chapter 2 pp 446–452
[43] Thorner D A, Willett P, Wright P M and Taylor R 1997 Similarity searching in files of three dimensional chemical structures: representation and searching of molecular electrostatic potentials using field-graphs ( Journal of Computer-Aided Molecular Design vol 11) chapter 96 pp 163–174
[44] Hahn M 1997Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds (Journal of Chemical Information and Modeling vol 37) chapter 96 pp 80–86
[45] Willett P, Finn and Morris G M 2013 WIRES Comp. Mol. Sci. vol 3 pp 226
[46] Maggiora G M, Vogt M, Stumpf D, and Bajorath J 2013 Molecular similarity in medicinal chemistry (J. Med. Chem. vol 57) chapter 8 pp 3186–3204
[47] M. A. Rabbih, S. M. A. Mamoun, Ezzat T. M. Selim, 2012, Semiempirical Calculations of the Proton Affinities of Some Quinolone Derivatives (Journal of Computer Science & Computational Mathematics vol. 2) pp. 53-55