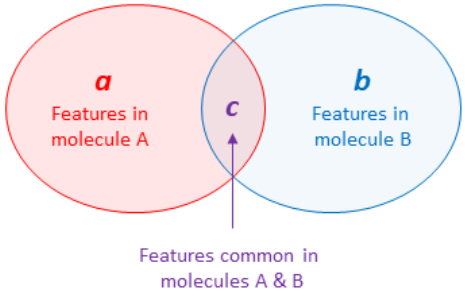
6.2: Similarity Coefficients

Many similarity metrics have been proposed and some commonly used metrics in cheminformatics are listed below, along with their mathematical definitions for binary features.



Metric name	Formula for binary variables	Minimum	Maximum
Tanimoto (Jaccard) coefficient	$S_{AB} = \frac{C}{A+B-C}$	0	1
Dice coefficient (Hodgkin index)	$S_{AB} = \frac{2C}{A+B}$	0	1
Cosine coefficient (Carbo index)	$S_{AB} = \frac{C}{\sqrt{ab}}$	0	1
Soergel distance	$D_{AB} = \frac{a+b-2c}{a+b-c}$	0	1
Euclidean distance	$D_{AB} = \sqrt{a+b-2c}$	0	N^{α}
Hamming (Manhattan or city-block) distance	$D_{AB} = a+b-2c$	0	N^{α}

$^{\alpha}$ The length of molecular fingerprints.

In the above table, the first three metrics (Tanimoto, Dice, and Cosine coefficients) are similarity metrics (S_{AB}), which evaluates how similar two molecules are to each other. The other three (Soergel, Euclidean, and Hamming coefficients) are distance or dissimilarity metrics (D_{AB}), which quantify how dissimilar the molecules are. These dissimilarity measures can be converted into similarity measures in a simple way. For example, for dissimilarity metrics whose possible values range from 0 to 1 (e.g., Soergel distance), the similarity score (S_{AB}) between two molecules can be computed simply by subtracting the dissimilarity score from unity:

$$S_{AB} = 1 - D_{AB}$$

Note that the Soergel distance between two molecules is the complement of their Tanimoto coefficient (that is, the sum of the two metrics is 1), while they are developed independently of each other.

If a distance metric has an upper-bound value greater than 1, (e.g., Euclidean or Hamming distance), the following equation [1] can be used to convert the dissimilarity score to the similarity score:

$$S_{AB} = \frac{1}{1+D_{AB}}$$

According to this equation, if two molecules are identical to each other, the distance (D_{AB}) between them is zero, and their similarity score (S_{AB}) becomes 1. On the other hand, as the D_{AB} value increases (i.e., for dissimilar molecules), the S_{AB} score approaches to 0.

An important question about molecular similarity evaluation is “how similar is similar?”. To answer this question, it is necessary to have a similarity threshold that can be used to determine whether molecules are similar enough. In 1996, Patterson et al. [2] analyzed sets of active compounds selected from scientific articles and showed that a Tanimoto coefficient of 0.85 or greater reflected a high probability of two compounds having the same activity. Since then, this Tanimoto value of 0.85 has been used in many studies as a general threshold for molecular similarity evaluation. However, as demonstrated in several studies [3], different molecular fingerprints give different similarity score distributions. For example, the Tanimoto score of 0.85 computed from MACCS keys have a different probability of the two compounds sharing the same activity than the probability represented by the same Tanimoto value (0.85) computed from ECFPs. The programming assignments for this chapter will help understand the impact of different molecular fingerprints upon computed similarity coefficient values.

References

1. Todeschini R, Ballabio D, Consonni V, Mauri A, Pavan M: **CAIMAN (Classification And Influence Matrix Analysis): A new approach to the classification based on leverage-scaled functions.** Chemometrics Intell Lab Syst 2007, 87:3-17.
2. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE: **Neighborhood behavior: A useful concept for validation of ''molecular diversity'' descriptors.** J Med Chem 1996, 39:3049-3059.
3. Jasial S, Hu Y, Vogt M, Bajorath J: **Activity-relevant similarity values for fingerprints and implications for similarity searching [version 2; peer review: 3 approved]**. F1000Research 2016, 5