Experimental evaluation of similarity indices about the MIA molecular family

Yassine M. Mejri[1,2] and Mehdi A. Beniddir[2] and Olivier Cailloux[1] and Meltem Öztürk[1]

[1]Université Paris Dauphine, PSL Research University, CNRS, Lamsade, 75016 Paris, France.
[2]Université Paris-Saclay Faculty of Pharmacy, CNRS, BioCIS, 91400 Orsay, France France.
mohamed-yassine.mejri@dauphine.eu, mehdi.beniddir@universite-paris-saclay.fr, olivier.cailloux@dauphine.fr, meltem.ozturk@dauphine.fr

Abstract. The purpose of this article is to evaluate the performance of the Morgan/Tanimoto, modified cosine similarity, Spec2Vec and MS2Deep Score similarity indices that aim at measuring structural similarities between molecules. We use a database containing about hundred of the emblematic monoterpene indole alkaloids (MIA) family of molecules.

1 Introduction

Comparing two molecules in order to see how similar they are is an important issue for analytical chemistry. Such comparisons are used in order to detect new molecules, to classify known molecules with respect to their utilities, their potential use, etc. Globally, two types of information may be used in order to compute similarities between molecules: information on their atomic structures or information coming from mass spectrometry.

There exist different way to compare such information. In this article we analyse the performance of four similarity indices (generally denoted by σ): one based on their atomic structure and three based on mass spectra.

- Structure based indices: we use the *Morgan/Tanimoto index* [Morgan, 1965] that we denote by σ_T.
- Mass spectra based indices: this category includes the rule based index called *modified cosine* (denoted by σ_C) and two machine learning based indices: *Spec2Vec*(σ_S)[Huber et al., 2021a] and *MS2Deep score* (σ_D) [Huber et al., 2021b].

We use these indices as implemented in the Python library MatchMS [Huber et al., 2020] and RDkit [Landrum, 2013].

Only a few articles analyse these indices in the literature (see for instance Bittremieux et al. [2022] and Huber et al. [2021b]). Our study is novel on two aspects: the type of molecules that we use in our experiments and the way that we evaluate the indices.

In our experiments we only use molecules belonging to the MIA (Monoterpene Indole Alkaloids) family. MIAs are known for their various biological activities and isomeric possibilities leading to analytical intricacies with respect to their annotation. To the best of our knowledge, our experiments are the only ones which are specially designed for this family.

In chemistry, there exists, for each molecular family, a classification of molecules depending on their atomic structures, called *skeletons classification*. For MIAs, chemists have defined 42 skeletons [Buckingham et al., 2010]. In our experiments, we compare the similarities provided by the four indices with the skeleton classification.

We denote by M the set of molecules in our database, by $n = |M| = 110$ their number, by S the set of skeletons and by $s : M \to S$ the function assigning each molecule to its skeleton. Our database spans $|s(M)| = 28$ of the 42 known MIA skeletons.

2 Similarity index vs skeleton

A similarity index is a function $\sigma : M \times M \to [0, 1]$ (the higher the number $\sigma(m_1, m_2)$, the most similar the index σ considers the molecules m_1, m_2). Apart from the four indices σ_T, σ_C, σ_S and σ_D introduced here above, we let $k = \mathbb{1}_{[s(m_1)=s(m_2)]}$, the "ground truth" index, denote the function that returns 1 if m_1 and m_2 have the same skeletons, 0 otherwise.

We define the average difference between two similarity indices σ_1 and σ_2 as

$$\epsilon_{\sigma_1,\sigma_2} = \frac{1}{\frac{n(n+1)}{2}} \sum_{i=1}^{n} \sum_{j=i}^{n} |\sigma_1(m_i, m_j) - \sigma_2(m_i, m_j)|.$$

We use this indicator in order to compare the four indices to the skeleton classification. We also analyse the difference between the structure based index σ_T with the three indices based on mass spectra. Table 1 shows our results.

σ	$\epsilon_{\sigma,k}$	$\epsilon_{\sigma,\sigma_T}$
σ_T	0.22	0
σ_C	0.25	0.18
σ_S	0.39	0.21
σ_D	0.72	0.54

Table 1: difference between index measures and skeleton classification and structure based index.

We see that the similarity index σ_T based on structure has the lowest average difference with the k index. This is expected as σ_T uses the atomic structure of the molecules, an information more directly relevant to the structure of the molecule than mass spectras. However, the three other indices are the ones that are generally useable in practice, as chemists who are trying to discover new molecules generally do not know their atomic structure (but may learn about their mass spectra thanks to specific devices). We also see that both machine learning based indices currently existing are further away from correct skeleton classification than the commonly used modified cosine index, an observation that suggests promising directions for future work.

One can observe that the computation of $\epsilon_{\sigma,k}$ for $\sigma \in \{\sigma_T, \sigma_C, \sigma_S, \sigma_D\}$, as displayed in the first column of table 1, mixes two kinds of indices since the codomain of σ is the interval $[0,1]$ while the codomain of k is the set $\{0,1\}$. In practice chemists often consider two molecules as similar if the similarity index is greater than 0.6. For this reason, we also use this threshold in order to discretize our similarity indices and thus define $\sigma_{0.6} = \mathbb{1}_{[\sigma(m_1,m_2)>0.6]}$, thus, $\sigma_{0.6}(m_i,m_j) = 1$ if $\sigma(m_i,m_j) > 0.6$ and $\sigma_{0.6}(m_i,m_j) = 0$ otherwise. Table 2 shows these supplementary results.

σ	True +	False +	False −	True −	$\epsilon_{\sigma_{0.6},k}$
σ_T	155	0	277	5673	0.045
σ_C	248	617	184	5056	0.13
σ_S	230	1082	202	4591	0.21
σ_D	410	5026	22	647	0.84

Table 2: Number of True positive, False positive, False negative and True negative comparisons. Note that $\epsilon_{\sigma_{0.6},k}$ is the proportion of (False positive + False negative).

We observe, again, that σ_T performs best and that σ_C achieve the best performance among the similarity indices that do not require atomic structural information.

3 Ongoing work

We are currently investigating clustering methods using the four similarity indices and comparing these clusters to the correct skeletons.

Another research direction that we are currently exploiting is the comparison of mass spectra based similarities with knowledge given by chemical experts. Instead of considering binary success or failure depending on whether a pair of molecules is correctly considered by some similarity index to belong to the same skeleton, a much richer information, and closer to actual practice, consists in considering how far away skeletons are from each other: different skeletons may be very close or very far away in terms of structure. To this respect, our chemist expert has specified distances between different skeletons. This information can then be used to further evaluate similarity indices, according to whether they tend to consider molecules that belong to similar skeletons as being similar.

4 Discussion

The starting point of our study was our desire to investigate the reliability of indices based on mass spectra. Such indices are systematically implemented in the discovery process of new molecules. During this analysis based mainly on mass spectrometry, information about a large number of molecules is obtained without knowing their atomic structure. For that reason, an automated similarity analysis is needed. Our current results suggest that modified cosine provides the most accurate predictions on the similarities that would be obtained if the atomic structure of molecules were known. It is also the one closest to the skeleton classification of MIAs. However, interpretation of the performance results shown here have to take into account the high proportion of negative pairs (5673 out of 6105) in our sample.

As future work, we will pursue the just described research paths. The second and most important step of our study will be the use of such index for the discovery of new MIA molecules.

REFERENCES

W. Bittremieux, R. Schmid, F. Huber, J. J. van der Hooft, M. Wang, and P. C. Dorrestein. Comparison of cosine, modified cosine, and neutral loss based spectral alignment for discovery of structurally related molecules. *bioRxiv*, 2022.

J. Buckingham, K. H. Baggaley, A. D. Roberts, and L. F. Szabo. *Dictionary of Alkaloids with CD-ROM*. CRC press, 2010.

F. Huber, J. J. van der Hooft, J. H. Spaaks, F. Diblen, S. Verhoeven, C. Geng, C. Meijer, S. Rogers, A. Belloum, H. Spreeuw, et al. matchms. 2020.

F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers, and J. J. Van Der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS computational biology*, 17(2):e1008724, 2021a.

F. Huber, S. van der Burg, J. J. van der Hooft, and L. Ridder. Ms2deepscore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of cheminformatics*, 13(1):1–14, 2021b.

G. Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2): 107–113, 1965.