

H. Xiang, J. Holliday and P. Willett

Chemoinformatics Research Group, Information School, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK.

Introduction

Any similarity measure that is to be used for similarity-based virtual screening (SBVS) has three principal components: the structure representation, the weighting scheme, and the similarity coefficient [1]. The many previous studies of SBVS that have been carried out have demonstrated that effective screening can be achieved using binary fingerprints and the Tanimoto coefficient.

A previous study [2,3] discussed the interactions between representation, weighting scheme and similarity coefficient when a chemical similarity measure is produced. In their study, it was shown that the Tanimoto coefficient is effective for binary (unweighted) similarity searching, but it has been suggested that other coefficients may be superior to the Tanimoto coefficient when applied to non-binary representations [4]. In this study, the cosine coefficient is compared with the Tanimoto coefficient in SBVS.

Methods

In this study, we adopted five kinds of weighting schemes to weight both the reference molecules and the molecules from two databases and applied two similarity coefficients, Tanimoto and cosine, to the resulting weighted fingerprints. The databases used were the MDL Drug Data Report (MDDR) and the World of Molecular Bioactivity database (WOMBAT).

The molecules were characterised by Scitegic ECFC_4 extended connectivity fingerprints, generated using Pipeline Pilot software.

The experimental process was as follows: ten active molecules were selected from each activity class as the reference molecules; the reference molecules and the database molecules were weighted using five different weighting schemes; the similarity coefficients were applied to a similarity search procedure and the performance of the different weighting schemes measured in terms of the number of active molecules retrieved in the top 1% of the database. For each activity class/weighting scheme/coefficient combination, the median performance measure for the ten searches was recorded. The median value was used to mitigate the effects of the classes that are tightly clustered in chemical space and that can retrieve very large numbers of actives.

Weighting Schemes

If $x(i)$ denotes the i th non-zero element in a fingerprint, f_i corresponds to the number of occurrences of the i th non-zero element, then the five weighting schemes are:

$$\begin{aligned} \text{W1: } x(i) &= 1 \\ \text{W2: } x(i) &= f_i \\ \text{W3: } x(i) &= \ln(f_i) \\ \text{W4: } x(i) &= \sqrt{f_i} \\ \text{W5: } x(i) &= 0.5 + 0.5 \frac{f_i}{\max\{f_i\}} \end{aligned}$$

In W5, $\max\{f_i\}$ stands for the largest value of f_i in a particular molecule's fingerprint.

Similarity coefficients

If we use x_i to represent the value of a certain element of the reference molecule and y_i for the value of the same element of a database molecule, then the similarity coefficients we applied are:

$$\text{Tanimoto: } S_{xy}(T) = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_i y_i}$$

$$\text{cosine: } S_{xy}(C) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \times \sum_{i=1}^n y_i^2}}$$

References

1. Willett, P. Similarity Methods in Chemoinformatics. *Ann. Review Inf. Sci. Technol.* **2009**, *43*, 3-71
2. Arif, S. M.; Holliday, J. D.; Willett, P. Analysis and Use of Fragment Occurrence Data in Similarity-Based Virtual Screening". *J. Comput.-Aided Mol. Design* **2009**, *23*, 655-668.
3. Arif, S. M.; Holliday, J. D.; Willett, P. Inverse Frequency Weighting of Fragments for Similarity-Based Virtual Screening. *J. Chem. Inf. Model.* **2010**, *50*, 1340-1349.
4. Al Khalifa, A., Haranczyck, M. & Holliday, J. Comparison of non-binary similarity coefficients for similarity searching, clustering and compound selection. *J. Chem. Inf. Model.* **2009**, *49*, 1193-1201.

Results

The notation Mab is used to identify different similarity measures, where a represents the weighting scheme that is applied to fingerprints of database molecules and b represents the weighting scheme that is applied to the fingerprint of a reference molecule. For example, M15 represents the similarity measure using W1 as the weighting scheme for a database molecules and W5 as the weighing scheme for a reference molecule.

Figures 1 and 2 compare the two coefficients over the 25 weighting combinations on the two databases. They indicate that, generally, the cosine coefficient retrieves greater numbers of active molecules than the Tanimoto coefficient. For symmetrical measures, i.e. those where the same weighting scheme has been used for the database and reference molecules, the two coefficients tend to perform comparably; whereas the Tanimoto seems to perform poorly where the measures are asymmetric. Performance is generally poor for measures which involve the weighting scheme W3, probably due to the fact that the log function produces zero elements for single fragment occurrences.

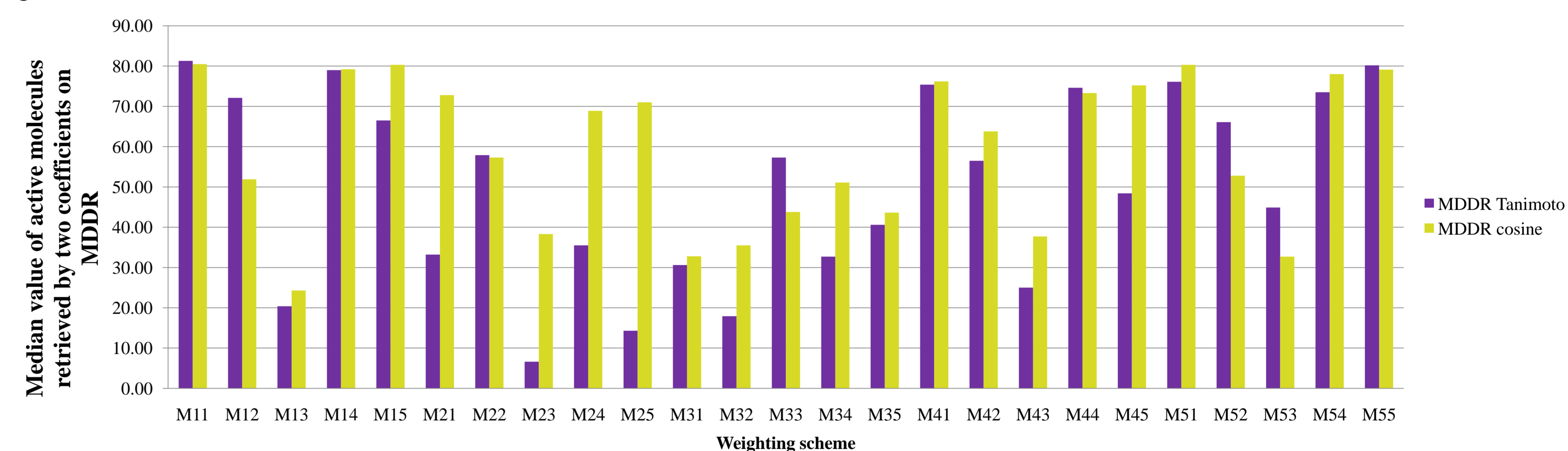


Figure.1 Comparison of two coefficients on MDDR

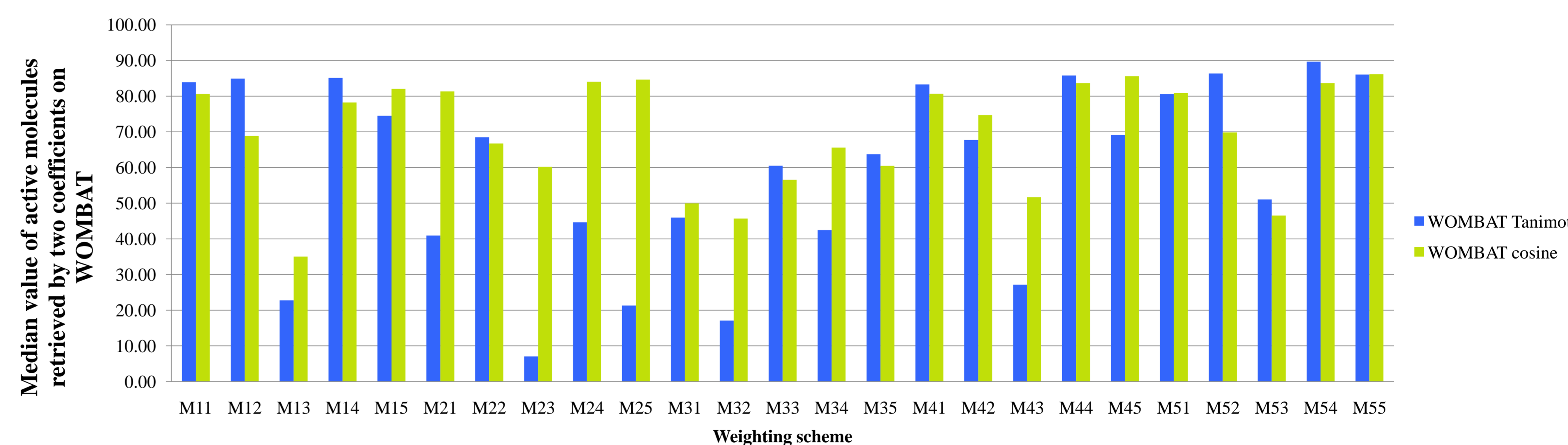


Figure.2 Comparison of two coefficients on WOMBAT

Conclusion

These experiments demonstrate clearly that, for asymmetric measures, the cosine coefficient shows better retrieval performance than the Tanimoto coefficient when applied to weighted occurrence data. Moreover, inspection of Figures 1 and 2 show that the cosine coefficient is noticeably less affected by changes in the nature of the weighting scheme that is used, whereas the Tanimoto coefficient shows reduced levels of performance with some types of weighting schemes.

Acknowledgments

Pipeline Pilot software and the MDL Drug Data Report database were provided by Accelrys, Inc.

The WOMBAT database was provided by Sunset Molecular.