Journal of Mathematical Chemistry 26 (1999) 263–265

Note

A proof of the triangle inequality for the Tanimoto distance

Alan H. Lipkus

Chemical Abstracts Service, P.O. Box 3012, Columbus, OH 43210-0012, USA

Received 20 April 1999; revised 3 September 1999

A distance, or dissimilarity measure, can be defined based on the Tanimoto coefficient, a similarity measure widely applied to chemical structures. A new, simple proof that this distance satisfies the triangle inequality is presented.

The clustering and similarity searching of large chemical structure files are well-established techniques [7,8]. Both require a sufficiently rapid method for calculating the similarity between two structures. The method most commonly used is to represent each structure by a vector in which every entry corresponds to some structural feature; the value of an entry is nonzero only if the corresponding feature is present in the structure. The value generally used is 1, in which case the vectors are binary (bit) vectors, but sometimes different values are used for different features to give greater weight to features deemed more important. The calculation of similarity between two structures is then a matter of quantifying, by some appropriate measure, the similarity between their respective vectors. The most widely used similarity measure for this purpose is the Tanimoto coefficient [6,8].

Consider a set of vectors of the form $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$, where x_{ik} is either 0 or w_k, a positive weight assigned to the kth entry (note that this weight does not depend upon i, meaning it is the same for all vectors). The Tanimoto coefficient for a pair of such vectors, $\mathbf{X}_m$ and $\mathbf{X}_n$, is

$$S_{mn} = \frac{X_{mn}}{X_{mm} + X_{nn} - X_{mn}}, \tag{1}$$

where $X_{ij} = \mathbf{X}_i \cdot \mathbf{X}_j$. The value of S_{mn} ranges from 0 to 1. When $w_k = 1$ for all k, $\mathbf{X}_m$ and $\mathbf{X}_n$ are bit vectors, and S_{mn} equals the number of bits "on" in both vectors divided by the number of bits "on" in either vector (for the case of bit vectors, S_{mn} is also known as the Jaccard coefficient [2,8]).

It is often useful to define a distance, or dissimilarity measure, based on the Tanimoto coefficient. The Tanimoto distance is $D_{mn} = 1 - S_{mn}$ (for bit vectors only, this quantity is identical to the so-called Soergel distance [2,8]). A significant mathematical property of the Tanimoto distance is that it satisfies the triangle, or metric, inequality, i.e.,

$$D_{ab} + D_{bc} \geqslant D_{ac} \tag{2}$$

for any three vectors, $\mathbf{X}_a$, $\mathbf{X}_b$, and $\mathbf{X}_c$, in which the kth entries are either 0 or w_k as previously described (equation (2) is *not* necessarily satisfied for three arbitrary vectors). The triangle inequality is satisfied by a number of other dissimilarity measures [2,8]. It ensures the desirable property that any two vectors having low dissimilarity to a third vector will have low dissimilarity to each other. It also can be used as the basis for heuristics to improve search efficiency [1,3,4]. A proof of equation (2) has been given by Späth [5]. Presented here is a simpler proof that proceeds by a totally different argument. This proof, unlike that due to Späth, does not demonstrate equation (2) directly but demonstrates instead the corresponding inequality for the Tanimoto similarities S_{ab}, S_{ac}, and S_{bc}.

Equation (2) is satisfied at once if $S_{ab} \leqslant S_{ac}$ or $S_{bc} \leqslant S_{ac}$ because these relations imply $D_{ab} \geqslant D_{ac}$ and $D_{bc} \geqslant D_{ac}$, respectively. It is thus necessary to prove equation (2) only for the case in which $S_{ab} > S_{ac}$ and $S_{bc} > S_{ac}$. A rearrangement of equation (1) that is useful in this proof is

$$X_{mn} = \frac{S_{mn}}{1 + S_{mn}}(X_{mm} + X_{nn}). \tag{3}$$

It can be seen that the inequality

$$(\mathbf{X}_b - \mathbf{X}_a) \cdot (\mathbf{X}_b - \mathbf{X}_c) \geqslant 0$$

or

$$X_{bb} - X_{bc} - X_{ab} + X_{ac} \geqslant 0 \tag{4}$$

must be true since the product of the kth entries of $\mathbf{X}_b - \mathbf{X}_a$ and $\mathbf{X}_b - \mathbf{X}_c$ equals either 0 or w_k^2. Equation (3) can be applied to equation (4) to give

$$\begin{aligned} &\left(1 - \frac{S_{ab}}{1 + S_{ab}} - \frac{S_{bc}}{1 + S_{bc}}\right) X_{bb} \\ &\quad \geqslant \left(\frac{S_{ab}}{1 + S_{ab}} - \frac{S_{ac}}{1 + S_{ac}}\right) X_{aa} + \left(\frac{S_{bc}}{1 + S_{bc}} - \frac{S_{ac}}{1 + S_{ac}}\right) X_{cc}. \end{aligned} \tag{5}$$

By applying equation (3) to the self-evident inequality $X_{aa} \geqslant X_{ab}$, it is found that $X_{aa} \geqslant S_{ab} X_{bb}$. It is thus valid to write

$$\left(\frac{S_{ab}}{1 + S_{ab}} - \frac{S_{ac}}{1 + S_{ac}}\right) X_{aa} \geqslant S_{ab}\left(\frac{S_{ab}}{1 + S_{ab}} - \frac{S_{ac}}{1 + S_{ac}}\right) X_{bb}, \tag{6}$$

since $S_{ab} - S_{ac} > 0$ by assumption. An analogous argument leads to

$$\left(\frac{S_{bc}}{1 + S_{bc}} - \frac{S_{ac}}{1 + S_{ac}}\right) X_{cc} \geqslant S_{bc}\left(\frac{S_{bc}}{1 + S_{bc}} - \frac{S_{ac}}{1 + S_{ac}}\right) X_{bb}. \tag{7}$$

Equations (5)–(7) imply that

$$\left(1 - \frac{S_{ab}}{1+S_{ab}} - \frac{S_{bc}}{1+S_{bc}}\right)X_{bb} \geqslant S_{ab}\left(\frac{S_{ab}}{1+S_{ab}} - \frac{S_{ac}}{1+S_{ac}}\right)X_{bb} + S_{bc}\left(\frac{S_{bc}}{1+S_{bc}} - \frac{S_{ac}}{1+S_{ac}}\right)X_{bb}. \tag{8}$$

It can be assumed that X_{bb} is not zero (equation (2) is automatically satisfied otherwise). Canceling X_{bb} in equation (8) and grouping terms with like denominator gives

$$1 \geqslant \frac{S_{ab} + S_{ab}^2}{1+S_{ab}} + \frac{S_{bc} + S_{bc}^2}{1+S_{bc}} - S_{ac}\left(\frac{S_{ab} + S_{bc}}{1+S_{ac}}\right)$$

or

$$1 + S_{ac} \geqslant S_{ab} + S_{bc}. \tag{9}$$

Equation (9) yields equation (2) when the Tanimoto similarities are expressed in terms of Tanimoto distances according to the relation $S_{mn} = 1 - D_{mn}$.

Acknowledgement

The author thanks Peter Willett for his helpful and encouraging comments on this work.

References

[1] W.A. Burkhard and R.M. Keller, Some approaches to best-match file searching, Comm. ACM 16 (1973) 230–236.
[2] J.C. Gower, Measures of similarity, dissimilarity, and distance, in: *Encyclopedia of Statistical Sciences*, Vol. 5, eds. S. Kotz and N.L. Johnson (Wiley-Interscience, 1985) pp. 397–405.
[3] M. Shapiro, The choice of reference points in best-match file searching, Comm. ACM 20 (1977) 339–343.
[4] D. Shasha and T.-L. Wang, New techniques for best-match retrieval, ACM Trans. Inf. Syst. 8 (1990) 140–158.
[5] H. Späth, *Cluster Analysis Algorithms* (Ellis Horwood, 1980) pp. 27–28.
[6] T.T. Tanimoto, An elementary mathematical theory of classification and prediction, IBM Report (November, 1958), cited in: G. Salton, *Automatic Information Organization and Retrieval* (McGraw-Hill, 1968) p. 238.
[7] P. Willett, *Similarity and Clustering in Chemical Information Systems* (Research Studies Press, 1987).
[8] P. Willett, J.M. Barnard and G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.