APPLICATION NOTE

On the Comparison of Different Tests for Identification of a Compound from its Mass Spectrum

Zeev B. Alfassi

Department of Nuclear Engineering, Ben Gurion University, Beer Sheva, Israel

It is shown that identification tests of different dimensions or dimensionless should not be evaluated (for their efficiency to identify molecules from their mass spectrum) by comparing the tests for a molecule itself (repeated measurements) with other molecules. This kind of tests must have similar dimensions (units). Another possibility is the comparison of tests on the basis of success of correct prediction for "unknown" molecules from a library of standards.   (J Am Soc Mass Spectrom 2003, 14, 262–264) © 2003 American Society for Mass Spectrometry

Recently, Wan et al. [1] compared two tests of significance for the purpose of differentiating between mass-spectra of different molecules (which in their case are several pentanucleotides), i.e., the similarity index test and the spectral contrast angle test. In order to evaluate which of the two significance tests is a better one, from the point of view of distinguishing between two different pentanucleotides, they calculated both the spectral contrast angle (θ) and the similarity index (SI) for repeated measurements of the same compound (defined as self-θ and self-SI) and compared them to θ and SI of two different pentanucleotides. Their method of selecting the better significance test is by calculating the ratio of the test value (either θ or SI) of the two different compounds to that of the self-value and preferring the one with the larger ratio. Their conclusion based on these ratios is that while both methods can distinguish between the different pentanucleotides, the "contrast angle test" is more sensitive.

Discussion

This kind of comparison done by Wan et al. [1] suffers from the disadvantage that they are comparing criteria that have different dimensions (θ has dimensions of degrees, while SI is a dimensionless term) and different ranges (θ can be from 0° to 90°, whereas the range of SI is from zero to 100). In order to see the artificiality of their comparison, let us take as the test value not the spectral contrast angle-θ, but rather the cosine of θ,

which is actually the value directly calculated from the two mass-spectra

$$\cos \theta = \Sigma x_i y_i / (\Sigma x_i^2 \Sigma y_i^2)^{0.5}$$

Where x_i and y_i are the various components (intensities or weighted intensities) of the two mass-spectra and the summation is done over all the components of the two mass-spectra. From their data of θ the self-$\cos \theta$ is 0.9989 and the range of $\cos \theta$ for various combinations of different pentanucleotides is 0.379 to 0.641. Thus leading to ratios of $\cos \theta$ to self-$\cos \theta$ in the range of 1.56 to 2.63, compared to the range of values of the ratios of SI given by them of 12.0 to 15.9. If this kind of comparison is correct, then this test show that the SI test is more sensitive than the test of cosine of the spectral contrast angle.

The inadequacy of their comparison can be shown also in the opposite way. Similarly to their mathematical operation on $\cos\theta$ to yield θ, we can suggest a manipulation on SI. Let us stick to their criterion of the spectral contrast angle θ, but in the same time let us choose a new similarity index defined by squaring the previous SI. Thus

$$SI_{new} = SI^2$$

$$SI_{new} = 10^4 \{\Sigma [(y_i - x_i)/(y_i + x_i)]^2\}/N$$

Where the sum is over all the N components of mass spectra x and y. Actually for our purpose the normalization factors 10^4 and N could be dropped, as they do not change the ratios. For this definition of SI_{new}, the ratios of the SI_{new} of two different pentanucleotides to the self-SI_{new} range between 144 and 253 much larger

than the ratio of the θ's which is in the range of 18.5 to 25.1.

Wan et al. [1] try to explain mathematically why the spectral contrast angle method is a better test, however their mathematics is wrong. They explained the advantage of the spectral contrast angle method for examining the difference between two almost equal spectra that are similar in most of their components and differ only by one or two components as due to the larger effect of multiplication of the components in the calculation of θ, rather than the subtraction done in the calculation of SI. However, as we have already shown, if we look on cos θ which is the real calculated value by the multiplication there is no advantage over SI. Their mathematics is wrong for two reasons: (1) The summation is done over large number of components and the change of one component makes this change in the summation almost negligible. (2) The operation of multiplication in the numerator is compensated by the denominator, which consists of the product of the lengths of the two vectors; while the length of each vector involve the sum of the squares (i.e., products) of its components. The real reason for the larger sensitivity of using θ is due to the sensitivity of the arccosine (cos^{-1}) function. Thus cos θ = 0.999 leads to θ = 2.56° while cos θ = 0.99 yields θ = 8.11°. Thus a factor of 1.01 in cos θ leads to a factor of 3.17 in θ itself.

It can be exemplify by comparing the two vectors (1,1,2,1) and (1,2,1,1) which differ in two components and the two vectors (1,1,2,1) and (1,1,1,1) where there is a change in only one component. The ratio of cos θ's for the two cases is 1.1 while the ratio of the SI's is 1.41. It is true that the ratio of θ's itself is 1.62 but the ratio of SI$_{new}$ is 2.0.

Thus it can be seen that by doing mathematical manipulation one can make each of these methods being the more sensitive one e.g., (1) using arccosine (cos^{-1}) that is more sensitive than cosine in this range (2) using x^2 that is more sensitive than x.

One can say that the conclusion from this discussion is that for comparing two methods they should be of the same units and of the same range as was done by Stein and Scott [2] whose three criteria for identification of spectra are all dimensionless and have the range of zero to one. However, Wan et al. could give the angle in units of radians that are dimensionless. The change of units will not change their conclusions, as the ratios remain the same, and will not change our criticism. The real criticism is not the different dimensionality but rather the absence of dimensionality that allows the mathematical manipulation in this kind of comparison. Even if both tests use dimensionless criteria the comparison is meaningless since we can change the power in one of the tests and by this, changes the ratio of two compounds to the self-value. The only comparison which can show which of the tests is more reliable is either both tests have dimension and the same one, or by finding what percentage of known test compounds hit the correct one as the first rank when compared to a library of spectra as was done by Rasmussen and

Isenhour [3], McLafferty and coworkers [4] and Stein and Scott [2]. It must be admitted that the uncertainty (relative standard deviation) is larger for the ratio of the SI than the ratio of cosθ and the mathematical manipulation will not change it. However; it is not clear what are the importance of this uncertainty. It should be pointed out that the main contribution to this uncertainty is the uncertainty in the value of self-SI. Comparing two different pentanucleotides the uncertainty in the value of SI is considerably smaller than of cosθ (for SI the uncertainties are 1.54%, 1.17% and 0.58% while for θ the uncertainties are 15.6%, 6.65% and 9.34%), which indicate on more accurate possibility to distinguish between different nucleotides through SI than through cosθ.

Similarity Index

In this paper, Gross and coworkers change the definition they made earlier [5] for the similarity index. Instead of dividing by the intensity of one of the mass spectra they divide by the arithmetic mean of both spectra; however, why they take the arithmetic mean and not the geometric or the harmonic one is unknown, when all have the same unit. Maybe a larger question is why they use subtraction in the numerator. Logically the name seems inappropriate, it cannot be a similarity index, if being larger means that the spectra are less similar. The appropriate name to their definition might be the dissimilarity index. When the index is zero then the dissimilarity is zero, which means that they are completely similar. Yet, still remains the question why to take the differences of the components and not their ratios. It seems that instead of inventing artificial similarity indexes we should go back to the measure known in statistics for more than hundred years, in order to correlate two series of numbers. In a similar way to the test of cos θ which check the parallelism of the two vectors, we can look on the components of the two spectra as two sets of variables and we have to check if one set is linearly dependent on the other one. The test for this in statistics is the correlation coefficient known also as Pearson's coefficient. Since we don't want negative values we will take its absolute value, which is the square root of the determination coefficient- R^2.

$$|r| = (R^2)^{0.5} = |\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y})|/$$
$$\{\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2\}^{0.5}$$

Where the summation is done over all the components of the two spectra. Actually this is very similar to the equation for cosθ except that this is done after transforming all the components to a coordinate system that is centered at $(\bar{x},\bar{y})$. Both cosθ and r are equal to unity for perfect matching. The less is the matching the less are both cosθ and r. Comparing cosθ and r for almost identical vectors it can be seen that r is more sensitive, i.e., further from unity. However, it does not say which

one of them will be better for search from the library. Thus, for example for the two close vectors (1,2,3,4,5) and (1,2.1,3,4,5) $\cos\theta$ is 0.999916 while Pearson's r is equal to 0.999643. To see it more clearly we should look on their deviation from unity, which is the criterion for dissimilarity. In the case of $\cos\theta$ it is $8.4\cdot10^{-5}$ while for r it is $3.57\cdot10^{-4}$. Another example is (1,2,3.1,4.1,5) and (1,2.1,3,4,5.2) where $\cos\theta$ is 0.999426 and Pearson's r is 0.99694, i.e. the dissimilarities are $5.74\cdot10^{-4}$ and $3.06\cdot10^{-3}$.

Another advantage of r on any arbitrarily defined similarity index is that its calculation is already one of the function of several spread sheets as for example in Excel it is the either the CORREL or PEARSON functions. These two functions are actually the same function, written by different but equal equations.

We perform similar tests for identification in order to identify the location of a radioactive point source in the lung using four gammas detectors [6] and comparing the 4-dimension vector of counts of supposedly unknown source with the appropriate vectors of 56 standard points. It was found that for 224 "unknown" measurements the $\cos\theta$ test hit correctly 182 cases (81.3%) while Pearson's r hit the correct answer only in 159 cases (71.07%). Thus while r is more sensitive to small changes in the vector components, $\cos\theta$ is a better test for identification.

An important advantage of the $\cos\theta$ test on the Euclidean distance test [2] is that it does not require normalization. If correct normalization is done (i.e. normalizing the vectors to have unit length) than $\cos\theta$ and the distance are equal tests since $d^2 = 2(1-\cos\theta)$. Stein and Scott [2] obtained different results for the $\cos\theta$ and Euclidean distance tests because they normalized differently their vectors.

References

1. Wan, K. X.; Vidavsky, I.; Gross, M. L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle. *J. Am. Soc. Mass. Spectrom.* **2002,** *13,* 85–88.
2. Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectra Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass. Spectrom.* **1994,** *5,* 859–866.
3. Rasmussen, G. T.; Isenhour, T. L. The Evaluation of Mass Spectral Search Algorithm. *J. Chem. Inf. Comput. Sci.* **1979,** *19,* 179–186.
4. McLafferty, F. W.; Stauffer, D. B.; Twiss-Brooks, A. B.; Loh, S. Y. An Enlarged Data base of Electron-Ionization Mass Spectra. *J. Am. Soc. Mass. Spectrom.* **1991,** *2,* 438–440.
5. Lay, J. O.; Gross, M. L.; Zwinselman, J. J.; Nibbering, N. M. M. A Field Ionization and Collision ally Activated Dissociation/ Charge Stripping Study of Some $[C_9H_{10}]^{+\cdot}$ Ions. *Org. Mass Spectrom.* **1983,** *18,* 16–21.
6. Pelled, O.; Alfassi, Z. B. to be published.

REPLY

[1]We wish to reply briefly to the criticisms made by Dr. Alfassi in the accompanying article. Dr. Alfassi ignored the intention of our Application Note. Our purpose was to provide and evaluate methods for comparing product-ion spectra obtained by MS/MS in ion-chemistry or structure studies. The goal was a strategy that could be easily implemented using standard programs (e.g., Excel) by an individual investigator who is occasionally faced with the requirement to compare spectra of an unknown and a reference. In the article, we described a modification of the similarity index (SI), which we had published some years ago (Reference 5 in the Alfassi article), and compared it to an established method (spectral contrast angle), using both real and simulated data. For these instances, there is no library of spectra available, unlike for EI spectra, and the requirement for the investigator to modify a library-search algorithm is too onerous. Furthermore, Stein and Scott [1] already compared these two approaches, and three others, in a library-search evaluation and found that the spectral contrast angle is superior. Thus, the criticism that we failed to test the two approaches by using a library of spectra misses the point.

Another criticism is the "artificiality" of the comparison of the two methods. This criticism can be leveled at most comparison schemes. Clearly one can amplify the differences in any comparison by a mathematical operation. Dr. Alfassi chose to take the cosine of the angle of the spectral contrast angle and diminish its ability to distinguish. We could have chosen to enhance it by raising it to some arbitrary power. Nevertheless, our test for effectiveness was a statistical evaluation, which was not done by Stein and Scott [1]. We showed that our data sets could be distinguished in a statistically significant way with one method (i.e., spectral contrast angle) and not the other.

Furthermore, the comparison was of ratios where each value was divided by the self-value, allowing us to compare dimensionless numbers. Thus, we do not understand Dr. Alfassi's comment that the "basic error is that we are comparing criteria with different dimensions."

Katty Wan
Ilan Vidavsky
Michael L. Gross
Department of Chemistry
Washington University
St. Louis, MO, USA

Reference

1. Stein, S. E.; Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 859–866.