

application note

Search Strategies for IR Spectra - Normalization and Euclidean Distance vs. First Derivative Algorithm

Bio-Rad Laboratories, Inc., Informatics Division, Philadelphia, PA 19102, USA

Bio-Rad's purpose in this note is to try to help you with your selection of algorithms and interpretation of your results. This Application Note will explain the basics of normalizing spectra and examine the difference between the Euclidean Distance, and the First Derivative search algorithms.

Normalization

Normalization is used to compensate for the differences in sample quantity (concentration and/or path length) between the database spectrum and the unknown spectrum. This allows a search algorithm to compare intensities of the data points. Spectra which have been stored in a database are always normalized over the spectral range of the database. When searching less than the full spectra range of the database, the spectra must be renormalized over the new range before an accurate comparison can be made. This is handled automatically by Bio-Rad's KnowItAll® search software and should be handled by other search software during the search.

There are two commonly used methods to normalize spectral data: the dot product normalization, which essentially normalizes the spectrum based on the total area under the curve and the scaling normalization, which normalizes the spectrum based on the height of the strongest peak.

Search algorithms in many search software programs normalize the spectrum when searching against a database. An understanding of how algorithms work is essential to achieve the normalization effect you want. The Euclidean search normalizes spectra by the dot product method. This is done by dividing each data point for both the reference and the unknown by the square root of the dot product of its spectrum.

The Euclidean Hit Quality Index (HQL)* is calculated by summing the square of the difference between each data pair. In other algorithms, spectra are normalized by the scaling method. This is done by subtracting a constant from each data point, making the minimum point equal to zero. The data is then multiplied by a constant to make the maximum data point equal to one absorbance unit.

Simply put, the Euclidean algorithms normalize spectra using all of the data points in the spectrum. Other algorithms normalize spectra using the maximum and minimum data points of the spectrum.

Selecting the Best Algorithm for your Unknown

The Euclidean Distance algorithm is generally a good first choice for spectral searching. Even if the unknown is not in the database, you will often find similar types of compounds and be able to classify the unknown. However, the results are adversely affected if the baseline of the spectrum is not at zero absorbance units. The algorithm compares each data point in the spectrum with the database, including baseline points. The algorithm interprets an offset or slope in the baseline to be a difference between the two spectra.

The First Derivative algorithm reduces this effect. Rather than comparing each point in the unknown with the database spectrum, the algorithm compares the difference between a pair of points in the unknown and the same pair of points in the database spectrum. This will eliminate any differences caused by a baseline offset and minimize differences caused by a sloping baseline.

Disadvantages to the First Derivative algorithm are that if the unknown is not in the database, the top hits may be very different from your unknown. Also, the First Derivative algorithm is less tolerant to slight peak shifts ($>2\text{ cm}^{-1}$) than the Euclidean algorithm.

Figure 1 illustrates how the Euclidean algorithm “sees” the two spectra. Both spectrum A (sloping baseline) and spectrum B (flat baseline) have been normalized by scaling. If these two spectra were compared using a Euclidean algorithm, the differences in the slope of the baselines will accumulate as the algorithm sums the squares of the differences. This accumulated difference will affect the HQL, possibly to the point that the software does not include the matching database spectrum in its top hits.

Figure 2 illustrates how the First Derivative algorithm “sees” the same two spectra. There is now almost no difference between the baselines of the two spectra. Now, when the algorithm sums the squares of the differences, there will be very little accumulated difference due to the baseline.

Summary

If your unknown spectrum has a flat baseline near zero absorbance units, then a Euclidean algorithm would be a good first choice. If your unknown has a sloping baseline, then either correct the baseline before searching or use the First Derivative algorithm.

The Euclidean Distance algorithm is essentially trying to match areas under the curve. It does not have any knowledge of what is spectroscopically significant. This leads to bands with a broad area like the OH stretch, being weighted very heavily in the search due to its large area, while a sharp band such as a C=N may be ignored because of its small area compared to the rest of the spectrum.



Figure 1. In spectrum B (flat baseline), the Euclidean Distance algorithm should yield good results. In spectrum A (sloping baseline), the First Derivative algorithm should yield good results.

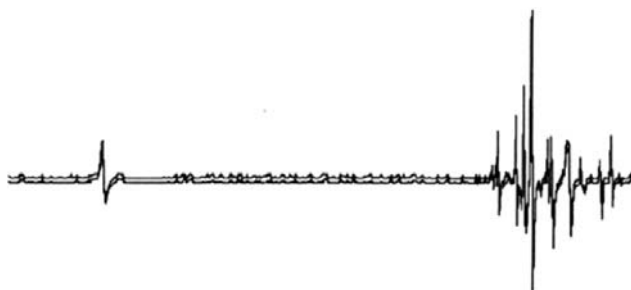


Figure 2. First Derivative of spectrum A overlaid on top of the First Derivative of spectrum B.

* Hit Quality Index (HQI) is a measure of how well an unknown spectrum matches a reference spectrum. Some spectral search software, including Bio-Rad's KnowItAll® software, report a high number, such as 999, to indicate a good match. Other search software may report a low number, such as 0.001, to indicate a good match.



**Bio-Rad
Laboratories**

Informatics Division
www.knowitall.com

China
Europe
Japan, Taiwan, Korea
Rest of World
USA

Phone: +1 267 322 6931 • E-mail: informatics.china@bio-rad.com
Phone: +44 20 8328 2555 • E-mail: informatics.europe@bio-rad.com
Phone: +81 03 (6361) 7080 • E-mail: informatics_jp@bio-rad.com
Phone: +1 267 322 6931 • E-mail: informatics.row@bio-rad.com
Phone: +1 267 322 6931 • 1 888 5 BIO-RAD (888-524-6723) • E-mail: Informatics.usa@bio-rad.com