application note

Examination of the Effect of Data Resolution on IR Spectral Database Search Results

Ty Abshear, Gregory Banik, Ph.D., and Marie Scandone
Bio-Rad Laboratories, Inc., Informatics Division
Two Penn Center Plaza, Suite 800, 1500 John F. Kennedy Blvd.
Philadelphia, PA 19102-1737 USA

Background

The pursuit of better search results has inspired innovation after innovation in the field of spectroscopy. It is commonly accepted that higher resolution of both reference databases and query spectra offer better comparison results for identification[1]. In the mid 1960's, Sadtler's original Spec-Finder product encoded up to 13 peaks at 11 intensity levels. Punch card limitations and computing power restricted options. Today, technological advances have changed disk space and computational time limitations of computer search methods even from the more recent past.

Storage space was a very significant concern only a decade ago. Hard disk drives averaged 10 GB or less, and storing and searching Sadtler's databases in high resolution would have been a strain. However, 4 to 8 GB of high-resolution data will not stress even the lowest end computer on the market today. The average storage cost per gigabyte has dropped from over $10 to less than $0.10 over the same period[2].

Computing power has also seen massive improvements. In the not too distant past, searching hundreds of thousands of spectra required nearly half an hour. Today's multi-core processors with a multi-threaded application like Bio-Rad's KnowItAll Informatics System[3] can perform the same task in only a few seconds. While storage cost has decreased by a factor of 100 in the last decade, the average computer's computational cost has decreased by a factor of more than 1,000.

Resolution

Confusion occasionally occurs when discussing resolution because the term is loosely interchanged with the data point spacing of a spectrum. Resolution defines the minimum distinguishable, closely spaced peaks. Since a peak requires three data points to be distinguished, the resolution is actually twice the data point spacing of the spectrum. In a spectrum with 0.96 cm^{-1} data point spacing, any peaks that are at least 1.92 cm^{-1} (its resolution) apart can be distinguished.

The relationship of resolution to valuable information that can be extracted from a spectrum is almost linear. There is a point of diminishing returns that may be different depending on the application. The difference between a 32 cm^{-1} data point spaced spectrum and a 1 cm^{-1} data point spaced spectrum is visually apparent, so the perceived value is significant. The differences between a 4 cm^{-1} data point-spaced spectrum and 1 cm^{-1} data point-spaced spectrum are not visually evident. However, mathematically the differences can be observed and the impact on search results traced.

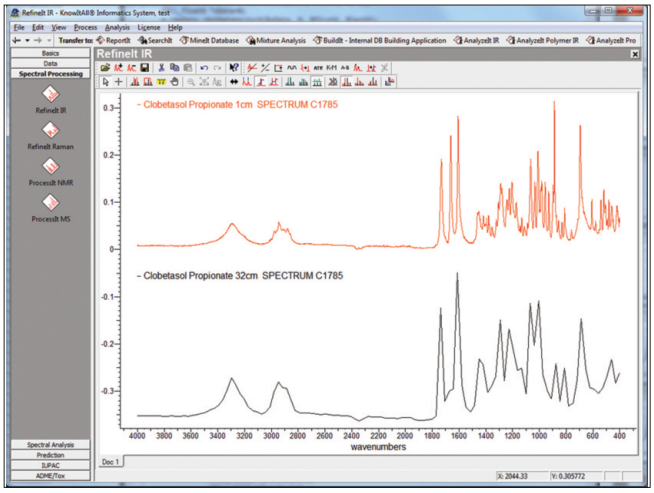

Figure 1. Clobetasol Propionate at 1 cm^{-1} and 32 cm^{-1}.

Materials and Methods

75,569 FT-IR spectra from Bio-Rad's Sadtler™ Standards Database[4] were analyzed at multiple resolutions for differences in curve area, dot product, peak loss, and peak shift. The original data point spacing of 0.96 cm^{-1} (with a 32 bit y-precision) was compared to linearly deresolved datasets at 2 cm^{-1}, 4 cm^{-1}, 8 cm^{-1}, 16 cm^{-1}, and 32 cm^{-1} data point spacing (with the same 32 bit y-precision).

The differences in curve area were compared by integration over the entire curve at different resolutions. Changes may be significant, even if unperceivable to the eye, because they

indicate smoothing effects over the entire spectrum. In sharper peak regions, deresolution tends to reduce intensity or cut off peaks altogether. In lower baseline regions, deresolution has less impact and tends to simply remove noise. However, spectra are typically normalized during a spectral search comparison. After normalization, sharper peak regions have increased area unless the peak has been removed (due to resolution data point changes), which offsets the area gain in other regions. This offsetting effect may mathematically hide larger changes in the spectrum's area.

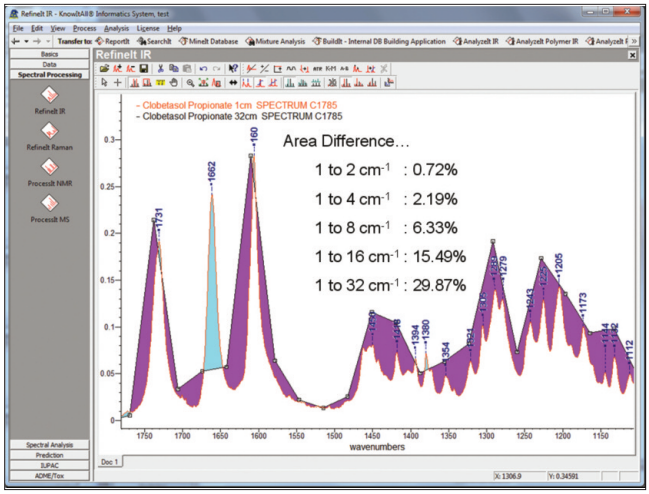

Figure 2. Clobetasol Propionate at 1 cm^{-1} and 32 cm^{-1} zoomed in with area differences highlighted.

Offsetting effects of the area make the dot product an even better comparison of differences in the spectrum's area at different resolutions. The dot product of a spectrum is its intensity squared and summed over its entire range. To compare dot products between spectra with different resolutions, the dot products must be normalized by dividing by the number of points in each spectrum. The dot product is significant because it is the main mathematical term used in every Euclidean Distance search algorithm. Differences in dot product will have a direct impact on Euclidean Distance search results.

Peak loss was identified using an automatic peak picking algorithm with a 2% noise threshold and an 8% global threshold. The total count of peaks in the original spectrum was then compared with the total count from the deresolved spectrum. Degrading the resolution results in peak loss and never in peak gain.

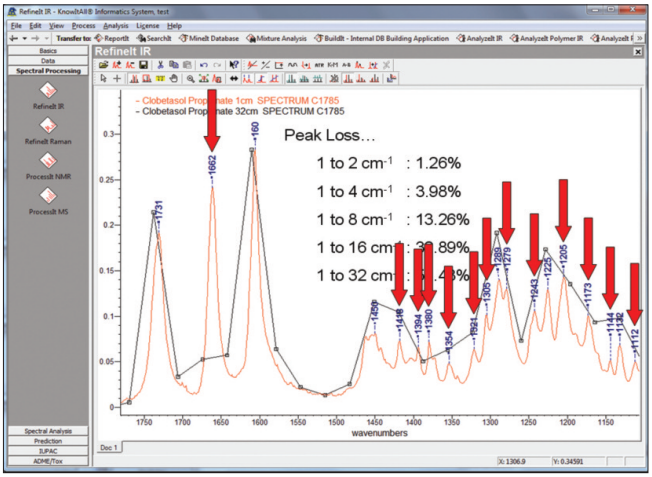

Figure 3. Clobetasol Propionate at 1 cm^{-1} and 32 cm^{-1} zoomed in with arrows indicating lost peaks.

Peak shift was analyzed one peak at a time after matching each peak from the original with the closest peak in the deresolved spectrum. The shift analysis only accounts for existing peaks and completely ignores peaks that have been lost in the deresolution process.

Additionally, an exhaustive statistical analysis of the top search results from 5,538 different query spectra was performed. Each query (none of which was present in the database) was searched against the same datasets with 0.96 cm^{-1}, 2 cm^{-1}, 4 cm^{-1}, 8 cm^{-1}, 16 cm^{-1}, and 32 cm^{-1} data point spacing three different times. A search was done using the Euclidean Distance, First Derivative Euclidean Distance, and Correlation search algorithms in Bio-Rad's KnowItAll Informatics System. The occurrence of changes in the order of the top three and the top ten search results were computed for each algorithm. In this analysis, results from 99,684 spectral searches—created from well over 7.5 billion spectral comparisons—were reviewed.

Results

The area, dot product, peak location, and number of peaks all change significantly after deresolution.

Table 1. Average percent change in area after deresolution.

Resolution Change	Area Difference
0.96 to 2 cm^{-1}	0.72%
0.96 to 4 cm^{-1}	2.19%
0.96 to 8 cm^{-1}	6.33%
0.96 to 16 cm^{-1}	15.51%
0.96 to 32 cm^{-1}	29.90%

Table 2. Average percent change in dot product after deresolution.

Resolution Change	Dot Product Change
0.96 to 2 cm^{-1}	1.39%
0.96 to 4 cm^{-1}	4.39%
0.96 to 8 cm^{-1}	12.68%
0.96 to 16 cm^{-1}	29.93%
0.96 to 32 cm^{-1}	54.61%

Table 3. Average percent change in peak loss after deresolution.

Resolution Change	Peak Loss
0.96 to 2 cm^{-1}	1.26%
0.96 to 4 cm^{-1}	3.99%
0.96 to 8 cm^{-1}	13.26%
0.96 to 16 cm^{-1}	32.26%
0.96 to 32 cm^{-1}	54.12%

Table 4. Average and percent peak shift relative to the resolution (twice the data point spacing) after deresolution.

Resolution Change	Average Peak Shift	Percent Peak Shift
0.96 to 2 cm^{-1}	0.73 cm^{-1}	18.25%
0.96 to 4 cm^{-1}	1.64 cm^{-1}	20.50%
0.96 to 8 cm^{-1}	3.86 cm^{-1}	24.13%
0.96 to 16 cm^{-1}	8.87 cm^{-1}	27.72%
0.96 to 32 cm^{-1}	18.40 cm^{-1}	28.75%

This data demonstrates an almost linear relationship between resolution and the mathematical information that can be utilized for identification. The largest impacts (steepest slopes) of resolution degradation appeared in the dot product and loss of peaks.

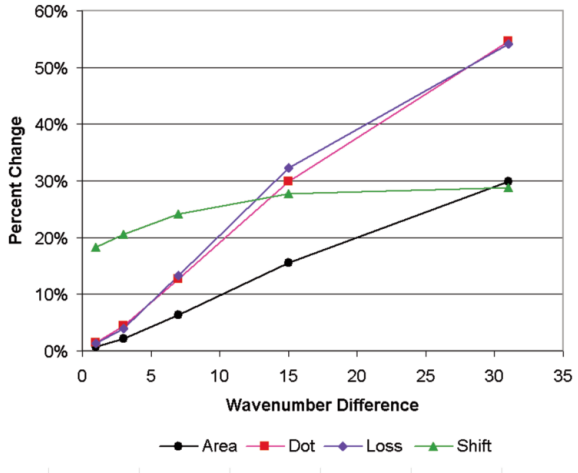

Figure 4. Plot of Area, Dot Product, Peak Loss, and Peak Shift changes after deresolution.

The exhaustive statistical analysis of 5,538 search queries found a non-linear impact on the search results.

Table 5. Percent change in search result order after deresolution using the Euclidean Algorithm.

Resolution Change	Change in Top Three Hits	Change in Top Ten Hits
0.96 to 2 cm^{-1}	2.9%	20.8%
0.96 to 4 cm^{-1}	7.5%	38.9%
0.96 to 8 cm^{-1}	32.1%	85.1%
0.96 to 16 cm^{-1}	70.7%	98.4%
0.96 to 32 cm^{-1}	92.7%	100.0%

Table 6. Percent change in search result order after deresolution using the First Derivative Euclidean Algorithm.

Resolution Change	Change in Top Three Hits	Change in Top Ten Hits
0.96 to 2 cm^{-1}	24.9%	84.2%
0.96 to 4 cm^{-1}	56.8%	97.7%
0.96 to 8 cm^{-1}	83.5%	99.9%
0.96 to 16 cm^{-1}	95.4%	99.9%
0.96 to 32 cm^{-1}	98.6%	100.0%

Table 7. Percent change in search result order after deresolution using the Correlation Algorithm.

Resolution Change	Change in Top Three Hits	Change in Top Ten Hits
0.96 to 2 cm^{-1}	2.5%	15.1%
0.96 to 4 cm^{-1}	6.3%	31.8%
0.96 to 8 cm^{-1}	31.7%	85.0%
0.96 to 16 cm^{-1}	71.7%	98.4%
0.96 to 32 cm^{-1}	93.1%	100.0%

The top three search results from all three comparison algorithms reveal the non-linear effect that resolution changes have on spectral search results. All three algorithms had changes over 30% of the time when degrading the data from its original 0.96 cm^{-1} to 8 cm^{-1} data point spacing. That is a resolution change of only 3.52 cm^{-1} .

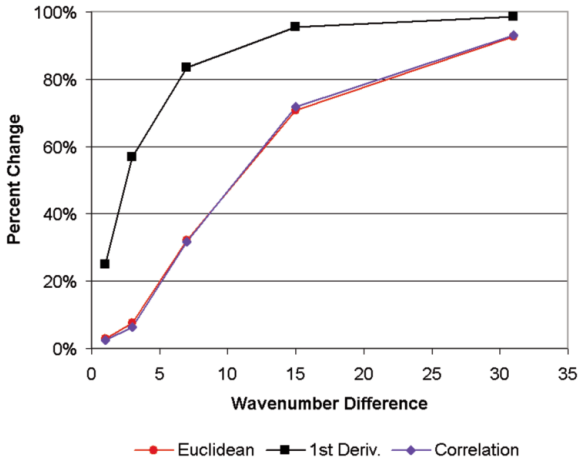

Figure 5. Percent change in the top three search results after deresolution.

The top ten search results accentuate the issue further with the steepest differences occurring with only minor changes to the resolution. The same resolution change of 3.52 cm^{-1} (from its original 0.96 cm^{-1} to 8 cm^{-1} data point spacing) results in search changes 85% of the time. Changes impacted the search results over 30% of the time degrading the data from its original 0.96 cm^{-1} to 4 cm^{-1} data point spacing. That is a resolution change of only 1.52 cm^{-1}.

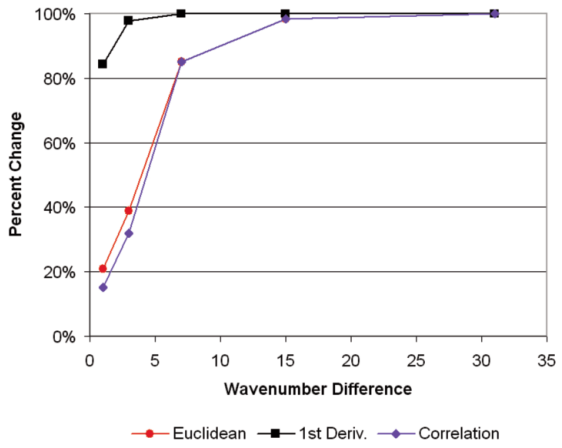

Figure 6. Percent change in the top ten search results after deresolution.

Conclusion

Altering an original spectrum's resolution has a significant impact on spectral comparison search results. Resolution changes do not need to be visually obvious to influence search results. Simply changing from 0.96 cm^{-1} to 4 cm^{-1} data point spacing will alter the top three results in as many as one out of every five searches.

There is a point of diminishing returns with respect to how high the resolution must be for any specific project. Changes in the top three search results are not significant when the correct compound is in the database. A well-run query spectrum can identify an exact match at very low resolutions with most comparison algorithms[5]. However, when a match is not in the database, interpreting results becomes an issue. Some software is designed only to display or analyze the top few results from spectral comparisons. Industry best practices recommend identifying separation in search result quality indices to aide in classification. Any resolution degradation clearly has some level of impact on quality index separation between search results.

Most alterations to a spectrum's range will also degrade the spectrum's original data points. The only exception is removing or padding data points without realigning them. Changing the range of a spectrum will typically re-space all of the data points within the spectrum and have a small smoothing effect that is often not visibly noticeable. The effect is directly proportional to the amount of data point shift relative to the data point spacing.

If possible, operations that may change the original data point spacing should be avoided. Special care should be taken when creating spectral databases in software that forces all records into the same range and resolution. Ideally, spectral search software should retain the original range and resolution of each database record. When searching, each individual database record should be compared with the query, and the resolution of the two spectra should be matched (always deresolving the higher resolution of the two spectra being compared and never artificially adding precision by interpolating data points in the lower resolution spectrum). Such a system avoids deresolution of spectra stored in a reference database as well as deresolution of query spectra to match a reference database with a resolution different from the query. Any data change should be approached with caution since it will have some degree of impact on search results whether or not the spectrum appears visually different.

References

(1) American Society for Testing and Materials (ASTM) International, ASTM Standard E 2310-04 (2009), Use of Spectral Searching by Curve Matching Algorithms with Data Recorded Using Mid-Infrared Spectroscopy, section 6.9.4.

(2) Komorowski, M. A History of Storage Cost. http://www.mkomo.com/cost-per-gigabyte (accessed February 2011).

(3) *KnowItAll® Informatics System*; version 8.3; Bio-Rad Laboratories, Inc., Informatics Division: Philadelphia, PA, 2010 (accessed February 2011).

(4) *Sadtler™ Condensed Phase IR Standards Database*; Bio-Rad Laboratories, Inc., Informatics Division, 2010 (accessed February 2011).

(5) American Society for Testing and Materials (ASTM) International, ASTM Standard E 2310-04 (2009), Use of Spectral Searching by Curve Matching Algorithms with Data Recorded Using Mid-Infrared Spectroscopy, section 6.10.1.

BIO-RAD

Bio-Rad Laboratories, Inc.

Informatics Division
www.knowitall.com

China	*Phone: +86 010 5939 0088 x381 • Email: informatics.china@bio-rad.com*
Europe, Middle East, Africa	*Phone: +44 20 8328 2555 • Email: informatics.europe@bio-rad.com*
India	*Phone: +91 124 4029300 • Email: informatics.india@bio-rad.com*
Japan, Taiwan, Korea	*Phone: +81 3 (6361) 7080 • Email: informatics_jp@bio-rad.com*
USA	*Phone: +1 267 322 6931 • Toll Free: +1 888 5 BIO-RAD (888-524-6723) • Email: informatics.usa@bio-rad.com*
All Other Countries	*Phone: +1 267 322 6931 • Email: informatics.worldwide@bio-rad.com*

280037-REV201207