White Paper

Analytical Methods for Field-Based Material Identification and Verification

Probabilistic Evaluation vs HQI Similarity Assessment

Key Words

Material identification, pharmaceutical screening, probabilistic approach, hit quality index, chemometrics, Thermo Scientific TruScan

Current field-based chemical identification instruments for pharmaceutical applications typically use one of three analytical methodologies: hit quality index (HQI), traditional chemometrics, or the Thermo Scientific probabilistic approach. This white paper compares HQI and the Thermo Scientific probabilistic approach, leaving the comparison between probability and traditional chemometrics to a separate white paper.

Hit Quality Index

Traditional methods for reference-library searching are typically based on the assessment of similarity metrics calculated via peak table comparisons, or more commonly, from those generated by full spectrum comparisons. Full spectrum approaches typically generate a “hit quality index” (HQI) between the unknown spectrum and each library spectrum. The HQI can be calculated based on Euclidean distance, median absolute deviation, or perhaps most frequently, the correlation coefficient between the test spectrum and each library spectrum. The correlation coefficient is equivalent to measuring the cosine of the angle between two spectra. The resulting correlation coefficient, R, is 1 when the two spectra are in perfect correspondence and 0 when they are orthogonal.

While a correlation coefficient threshold of 0.95 is frequently used to determine whether two spectra are a match, the correlation is merely an angle and not a probability. Thus, the traditional threshold of 0.95 in no way means 95% likelihood, 95% confidence, or 95% agreement. Furthermore, a correlation coefficient other than 0 or 1 has no direct interpretation in the context of spectral identity because a transparent interpretation as a test statistic only holds when dealing with random normal variates, clearly not the case for Infrared or Raman spectra. While the correlation coefficient has been a popular choice for pure material assessment, it is not particularly sensitive to discrepancies between spectra of interest.

Probabilistic Evaluation

As technical advances brought laboratory-quality instruments to the field, a new testing approach was needed to address the challenge of unknown chemical identification. In the process of identifying substances within a vast unknown library, handheld instruments put the power of spectroscopy into the hands of a new user – field technicians without extensive spectroscopy and chemical training. While HQI met the initial need for laboratory use, a new approach was required for these less experienced users who operate in challenging environments and sampling conditions.

Thermo SCIENTIFIC

An alternative to correlation-based library searches and a development-intensive classification method that has seen increased adoption in recent years is the comparison of measured data to library spectra in a probabilistic fashion. The probabilistic approach has been used on Thermo Scientific™ handheld Raman and Infrared devices since their inception.

Thermo Scientific™ TruScan RM™ and TruScan GP™ spectroscopic analyzers employ the probabilistic approach.

In the case of pure material evaluation, this procedure determines whether the measured spectrum of the unknown sample lies within the multivariate domain of a reference spectrum of interest. The multivariate domain is defined by the uncertainty characteristics of each measurement, which include measurement settings (e.g. exposure time and number of scans or sweeps), environmental properties (e.g. temperature, dark current) and the properties of the sample of itself (e.g. Raman cross section, absorbance, refractive index, etc.). When comparing spectra in the manner described above, the algorithm looks for features that contradict the reference model rather than determining how similar two spectra are (i.e. correlation with HQI).

Like most statistical tests, the analysis is distilled into a p-value, in this case the probability that the observed differences between the test and reference model simply arose by chance, given the uncertainty of the measurement. In statistical significance testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. As a common practice in statistics, the null hypothesis is rejected when the p-value is less than a certain significance level, often 0.05. This indicates that the observed result would be highly unlikely under the null hypothesis. In other words, the observation is highly unlikely to be the result of random chance alone. The null hypothesis in this context claims that a measurement spectrum belongs to the population of the reference library spectrum, given the measurement uncertainty. The alternative hypothesis claims that a measurement spectrum does not belong to the population of the reference library spectrum. Thus, p-value is the probability of observing a spectrum more extreme (worse) than the sample spectrum, if the sample spectrum belongs to the population of library spectrum (i.e. when null hypothesis is true).

To illustrate the effectiveness of the probabilistic approach, we consider the probabilistic comparison of Microcrystalline Cellulose to other celluloses. We test the null hypothesis (H0 = Microcrystalline Cellulose), the alternative (H_1 = not Microcrystalline Cellulose) and compare with the HQI result. Table 1 shows the p-value versus corresponding HQI values for Microcrystalline Cellulose, as well as corresponding results for three other celluloses, based upon second-order fluorescence baseline correction.

Table 1. In an evaluation of celluloses, the p-value approach correctly identifies the sample as different from the library reference (Microcrystalline Cellulose) while HQI does not.

	p-value	HQI
Microcrystalline Cellulose	0.338	0.9998
Hydroxyethyl Cellulose	0.00000754	0.9970
Methyl Cellulose	0.00000185	0.9766
Hydroxypropyl Cellulose	0.0000000323	0.9796

As values in Table 1 illustrate, the probabilistic and HQI approaches both correctly identify Microcrystalline Cellulose. In the probabilistic approach, we accept the null hypothesis (p-value > 0.05) and, in the HQI approach, the correlation coefficient is very nearly 1.0. However, in regard to the other celluloses, the probabilistic approach rejects the null hypothesis (p-value < 0.05) while the correlation method suggests reference matches with these materials – clearly returning false-positive results.

In Figure 1, we can visually examine the measured spectrum of Hydroxyethyl Cellulose (the reference sample) versus the Microcrystalline Cellulose library spectrum. The p-value for the Hydroxyethyl Cellulose sample spectrum, shown in Table 1, is 0.00000754. The p-value result below 0.05 indicates a low probability that the measured spectrum of the unknown sample lies within the multivariate domain of a reference spectrum of interest, if the null hypothesis is true (i.e. the unknown and library are different). Meanwhile, the HQI result of 0.9970 is a high correlation score (e.g. cosine angle), far above the traditional 0.95 passing threshold, yet the unknown material is Hydroxyethyl Cellulose, not Microcrystalline Cellulose.

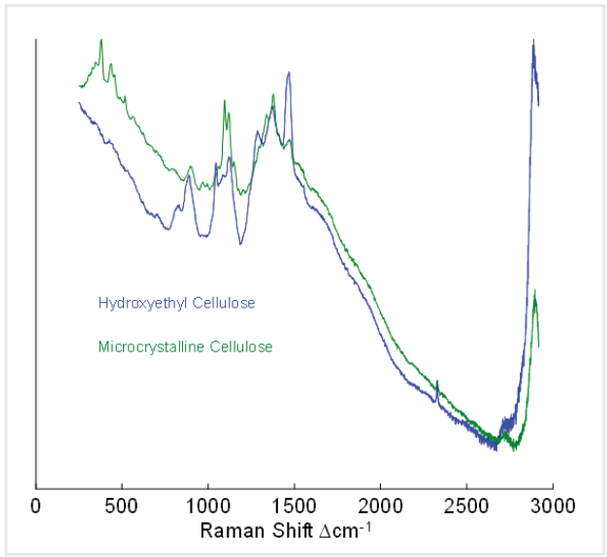

Figure 1. The measured spectrum of Microcrystalline Cellulose is visually different from the spectrum for Hydroxyethyl Cellulose, a difference confirmed by a p-value < 0.05.

To further illustrate the effectiveness of the probabilistic approach, we will consider the probabilistic comparison of 15% DMMP in chloroform to pure chloroform, as shown in Figure 2. In this case, we examine the unknown measured spectrum and the pure chloroform library spectrum, testing the null hypothesis (H_0 = pure chloroform) and the alternative (H_1 = not pure chloroform). With this assessment, it becomes very clear that the discrepancy in the 715 cm^{-1} region cannot be due to noise alone. The p-value is the probability of observing the unknown spectrum or one more extreme, if the null hypothesis is true. For this comparison, the value is calculated as 6.1 x 10^{-4}. Thus, if the sample were pure chloroform, the probability of observing a spectrum as extreme as the unknown measurement would be ~ 1 in 1639 – highly unlikely. Correspondingly, the algorithm would recognize that the sample cannot be pure chloroform, returning a p-value less than 0.05 (i.e. statistically significant).

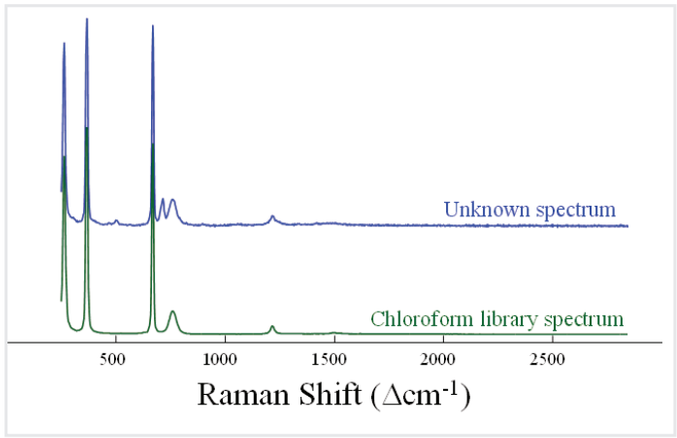

Figure 2. Comparison of 15% DMMP in chloroform to pure chloroform further illustrates the effectiveness of the probabilistic approach.

Summary

Both HQI and probabilistic methods are proven analytical techniques for interpretation of spectroscopic data. While HQI is well-suited for laboratory use by spectroscopy experts – its original and intended purpose – probabilistic analysis is specifically designed for field-based decision making, with very high accuracy. When considering these options, users should evaluate the simplicity and reliability of results in relation to relatively inexperienced users who operate in challenging environments and sampling conditions.

References

1. JA Swets et al., “Assessment of diagnostic technologies” Science 205:753-759 (1979)
2. TD Wickens, Elementary Signal Detection Theory, Oxford University Press 2001
3. GM Pesyna et al. “Probability based matching system using a large collection of reference mass spectra”
4. Analytical Chemistry 48:1362-1368 (1976)
5. Proc. of SPIE, spie.org, Vol. 6378 637809,1-11
6. CD Brown and RL Green, “Considerations for embedded authentication using intelligent portable Raman systems,” FACSS 2006, Orlando, FL
7. CD Brown, “Revisiting the Assumptions of Spectral Library Search in Light of New Frontiers,” FACSS 2005, Quebec, PQ
8. J Li, DB Hibbert, S Fuller, G Vaughn, “A comparative study of point-to-point algorithms for matching spectra,” Chemom. Intell. Lab. Syst. 82:50 (2006)
9. RL McCreery, Raman Spectroscopy for Chemical Analysis, Wiley (2000)
10. DL Massart et al., Handbook of Chemometrics and Qualimetrics: Part A, Elsevier (1997)

thermoscientific.com

Africa-Other +27 11 570 1840
Australia +61 2 8844 9500
Austria +43 1 333 50 34 0
Belgium +32 53 73 42 41
Canada +1 800 530 8447
China +86 10 8419 3588
Denmark +45 70 23 62 60
Europe-Other +43 1 333 50 34 0
Finland /Norway/Sweden +46 8 556 468 00
France +33 1 60 92 48 00
Germany +49 6103 408 1014
India +91 22 6742 9434
Italy +39 02 950 591
Japan +81 45 453 9100
Latin America +1 608 276 5659
Middle East +43 1 333 50 34 0
Netherlands +31 76 579 55 55
South Africa +27 11 570 1840
Spain +34 914 845 965
Switzerland +41 61 716 77 00
UK +44 1442 233555
USA +1 800 532 4752

Thermo SCIENTIFIC

Part of Thermo Fisher Scientific