X-ray Powder Diffraction Metrics

Authors:
George Runger, PhD
Affiliation - Arizona State University
Address - Department of Industrial Engineering
Arizona State University
PO Box 875906
Tempe, AZ 85287-5906

Kelly Canter, PhD
Affiliation -Pfizer Inc.
Address - 520/2194
2800 Plymouth Rd
Ann Arbor, MI 48105

John Twist, PhD
Affiliation - Mylan Laboratory
Address - 1500 Corporate Drive
Suite 400
Canonsburg, PA 15317

David Rossi, PhD
Affiliation - Mylan Laboratory
Address - 1500 Corporate Drive
Suite 400
Canonsburg, PA 15317

Clifford Kirkham
Affiliation – Arizona State University
Address - Department of Industrial Engineering
Arizona State University
PO Box 875906
Tempe, AZ 85287-5906

Keywords

X-Ray Powder Diffraction, D-line, and Difference Distribution

Abstract

To compare sample data from x-ray diffraction a suitable metric is developed to evaluate the similarity (or dissimilarity) between samples and to compare it to a references. The methodology focuses on the d-lines and uses the intensity measurements indirectly in the algorithm. The resulting d-line distribution and summary statistics of the differences is used to establish whether a true difference exist between a reference and sample material. The application of the methodology was then applied to real XRPD data. The methodology was able to correctly distinguish each of the samples from the published reference source. Further refinement of the methodology and future research is highlighted to further tolerate peak shifts.

1. Introduction

In the field of pharmaceutical sciences, crystal polymorphs represent an important area of study. The polymorphic form, that is the particular arrangement of atoms of the drug molecule in the crystal lattice, imparts important properties such as ease and rate of solubility or degree of stability to the drug substance. These features are important in that they can impact both the fraction of the dose absorbed by the patient and the overall chemical stability of the drug.

Although there are several instrumental techniques for characterizing polymorphs, the most widely used is that of x-ray powder diffraction (XRPD). In this technique, a milligram sample of the powdered chemical material to be characterized is placed onto a small rotating metal dish. The sample is then exposed to electromagnetic radiation at a standard x-ray wavelength. As the angle (the so-called 2-theta angle) of the incident radiation is varied, the test material will diffract the radiation to a greater or lesser extent. The intensity of the diffracted light is detected and is used to generate a characteristic diffraction pattern or diffractogram. Each polymorphic form is a different crystal construct and will diffract light slightly differently, so the peaks or lines of this diffractogram and the 2-theta angles at which they are produced are highly characteristic of the polymorphic form of the test material and can generally be used to identify and compare different polymorphic forms. Thus, an understanding and use of this technique has significant implications for regulatory and intellectual property (IP) issues.

Often a drug company will seek to obtain one or more patents, which claim selected polymorphic forms and the processes for making them, so that they can maintain a competitive advantage after the main composition-of-matter-patent for the drug expires. This is the fundamental principle of why patent laws exist, to support and encourage medical innovations. Other drug companies wishing to manufacture and sell the same drug will have to use unpatented or un-patentable polymorphic forms. For this reason, and to ensure the protection of IP, the ability to identify and compare the polymorphic form of pharmaceutical material of multiple batches has become an important legal question.

Before one can develop a scientific answer to this legal question, a variety of aspects to characterize polymorphs via XRPD must be addressed. One difficulty associated with this type of characterization is that the diffraction efficiency is highly dependent on the size and orientation of the crystals. For this reason, the dish containing the sample is slowly rotated to average out any differences in crystal orientation. While this helps to minimize the differences of crystal orientation, it is because of these crystal orientation differences that diffractograms often show dramatic differences in 2-theta line intensity. The line intensities are, therefore, not highly characteristic of polymorphic identity. While line intensity is helpful and merits some attention, it is not as a defining characteristic of polymorphic identity as 2-theta line position.

A typical comparison of two polymorphic forms would traditionally be done by simple visual comparison of their diffractograms, with emphasis being given to 2-theta line position over 2-theta line intensity. If all line positions of the two diffractograms agreed to within some tolerance ($\pm 0.2^{o}$ is a typical rule-of-thumb value) the polymorphic forms of the two samples are thought to agree. If one or more of the lines disagree then the level of confidence in a positive comparison is lower. If one or more lines from a sample are not observed in a second sample, but a number of major lines agree well then it is possible that the second sample is a different polymorph or a mixture of two polymorphic forms. This represents a second challenge in polymorph identification. Finally, because the original material is not always available and only literature representations of XRPD data may exist, from a practical point of view, any statistically based comparison approach would need to be able differentiate data obtained from different sources such as diffractograms published in the patent literature in comparison to diffractograms of materials generated in one's own lab.

Although visual comparison of diffractograms is the traditional approach, it has a number of problems, including a high reliance on the expertise of the individual doing the comparison, inability to automate the process, and hand manipulation of data in what can potentially be a data-intensive process. Additionally visual comparisons can sometimes use subjective and arbitrary rules of thumb, as described above, and it is possible to overlook important diffraction features, as would be the case with mixtures of polymorphs. A

variety of science-based approaches have been proposed for addressing the comparison issue. The most promising are computer based search/match programs that incorporate a series of algorithms iteratively to match up spectral data in powder diffraction databases [1, 2…]. This approach is effective but often can yield several unanticipated results. For example, spectral data from a sample may match up with a variety of different reference files contained within a database, due to the variety of excipient's used within a formulation. This is expected given the computational complexity and potential permutations of the d-lines that exist for an active index.

Other computer-based approaches involve using Artificial Neural Networks (ANN) or optimization algorithms [3, 4]. These techniques are more for recognizing and quantifying the physical features of the specimen under investigation, such as a bulk drug for example, but fail in establishing equivalence definitively. Statistical applications are a means for establishing equivalence, however most approaches to date utilize partial least squares (PLS), principle component analysis, or clustering techniques on spectral measurement data [5]. These applications try to cluster curves and match the best candidate from a reference set using a multivariate distance approach from the intensity measurements. More so, these applications have proven to be effective in quantifying noise in spectral data when applying designed experiments for process control [6]. Other statistical approaches have made use of Reitveld's method for refinement of spectral data, and then model it as a Poisson distributed random variable [7].

All these are effective means for comparison; they primarily focus on the fit of the spectral data to a known specimen and use the intensity data directly even though it is not the most defining characteristic. However, without quantifying the error in the spectra data, a true difference between a reference and sample diffractogram will be difficult to assess. This paper outlines a novel methodology for creating a score or metric that is capable of measuring the difference between samples. The method focuses on the d-lines and only uses the intensity measurements indirectly in the algorithm. It is based on computations that would not have been possible without the recent advances in computer technology. The application of the methodology is then applied to real XRPD data. Final comments are noted, as well as future research needs for further development.

2. Methodology

In order to compare X-ray diffraction data samples to a reference a method is needed to evaluate the d-line data. There are several important characteristics of such data. Because d-lines are obtained from intensity peaks, a higher or lower threshold applied to intensity can reduce or increase the number of d-lines, respectively. Consequently, the number of d-lines in a sample and a reference from the same material are not expected to match. Instead, a method is needed that can accommodate different numbers of d-lines, but still distinguish d-line patterns that are close to ones that are not.

Each sample provides a vector of n d-lines denoted as

$$d_s$$

This vector can be transformed to a vector of n – 1 differences as

$$\Delta d_s$$

This vector contains the information that is used to compare samples. Typically, a sample contains more d-lines than a reference from the same material. This is expected to be due to extra, extraneous peak in the intensity signal. The extra d0lines make it difficult to naively compare difference vectors between the sample and reference. Suppose the reference sample contain m d-lines denoted as the vector d_r. There are m – 1 differences and the vector of differences is denoted as Δd_r. The objective is to compare Δd_s with Δd_r.

The approach to be used makes use of the fact that samples from the same material should have some (or many) d-line difference that approximately match, but that a sample may have extraneous d-lines. Vectors d_r and d_s can be compared with an algorithm that selects m d-lines from the n in d_s to obtain a measure of similarity. Let the vector of m d-lines selected from d_s be denoted as

$$d_s^m$$

Also, let

$$\Delta d_s^m$$

denote the differences. The match is considered good if Δd_r is similar to Δd_s^m. Any number of distance metrics between vectors can be used. For example, Euclidean distance between the two *m*-dimensional vectors is reasonable.

However, a metric such as Euclidean distance is dominated by the larger differences. To provide approximately equal weight to all differences, a proportional difference is used. We measure the distance between a sample and reference as follows. Let the ith elements of a vector d_s be denoted as $d_s(i)$. The distance between d_r and the transformed vector d_s^m is measured by

$$D = \frac{1}{m}\sum_{i=1}^{m}\frac{\left|\Delta d_r(i) - \Delta d_s^m(i)\right|}{\Delta d_r(i)}$$

The division by *m* is used to account fro the fact that references also have different number of d-lines and the objective is to normalize the similarity metric for the number of d-lines in the reference.

The key is that the distance *D* is calculated for each possible set of d-lines in d_s^m. This adjusts for extra, or extraneous d-lines, and provides a distribution of distances. Finally, this distribution of distances can be summarized by its average, or other summary statistics, in order to evaluate the similarity of a sample to a reference vector of d-lines. This is discussed further in the following section.

3. Computations

If the reference contains *m* d-lines and the sample contains *n*, with $n \geq m$ the distance can be calculated fro every set of *m* d-lines selected from *n*. The number of possible sets is

$$n(S) = \frac{n!}{m!(n-m)!}$$

and this is feasible to calculate through complete enumeration for moderate values of *n* and *m*. For example, if $n = 20$ and $m = 15$ the number of possible sets is approximately 15,000. For every set *S* the distance *D* is calculated for that set and denoted as $D(S)$.

There are several reasonable measures of similarity between d_r and d_s. One is the minimum value obtained over the sets

$$(d_r, d_s) = \min_s D(S)$$

Another is the average value of D over the sets

$$(d_r, d_s) = \frac{\sum_s D(S)}{n(S)}$$

If the average of D(S) over the permutations is close to the reference then the sample is judged to be similar to the reference. A number of summaries of the distribution can be used to compare the sample to the reference. The distribution characterizes the relationship between the sample and reference in a manner that is robust to extraneous or missing d-lines. Other summaries such as the variance or median of the distribution can also be used to assess the similarity.

Upon review of the data and the output results it was noted that it might be useful to extract the d-lines from a sample at selected intensity thresholds. That is, the analysis might be done several times, each with a different intensity threshold to generate different sample d-lines. The threshold that produces the best match to the sample data might be used. However, some reasonable decisions for thresholds need to be made. With too high a threshold the sample data can collapse to only two or fewer d-lines and the ability to compare to a reference is limited. For these reasons an additional weighting based on intensity was used as a scalar for the distance between d_r and the transformed vector d_s^m, *D*. The difference equation and corresponding weights are as follows

$$D = \frac{1}{m}\sum(\Delta d_r(i) - \Delta d_s^m(i))^2 (w_s(i))^2 (w_r(i))^2$$

$$D = \frac{1}{m}\sum([d_r(i+1) - d_r(i)] - [d_s(i+1) - d_s(i)])^2 (w_s(i))^2 (w_r(i))^2$$

$$D = \frac{1}{m}\sum([d_r(i+1) - d_r(i)] - [d_s(i+1) - d_s(i)])^2 (\min(I_s(i+1), I_s(i)))^2 (\min(I_r(i+1), I_r(i)))^2$$

where $\Delta d_r(i)$ is the difference between two adjacent d-lines in the reference vector , $\Delta d_s(i)$ is the difference between two adjacent d-lines in the sample vector, $w_r(i)$ is the minimum intensity of two adjacent d-lines in the reference vector, and $w_s(i)$ is the minimum intensity of two adjacent d-lines in the sample vector.

This has the advantage that major d-lines are given more weight, while a focus is maintained on d-lines as the defining characteristic. Some experiments were completed for these metrics in order to evaluate the ability to discriminate between samples and references. In the following section an example with such a weighted d-line metric is shown.

4. Application

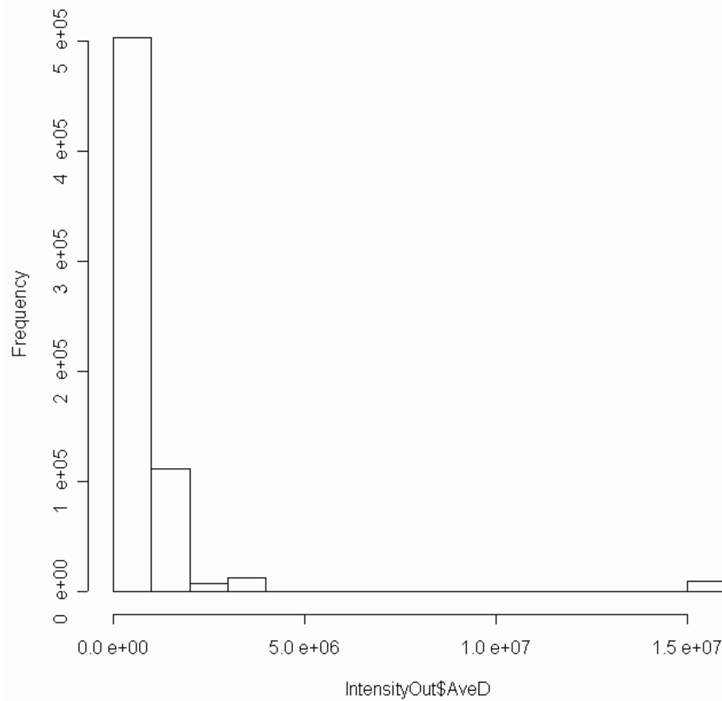
Samples were obtained from a single production lot, whereas the reference materials were obtained from a published referenced source. The actual d-line values for the reference were re-created via digitization software called "Ungraph it." This application takes a scanned image of a printed graph and converts it into a series of data based on user inputs.
There were three reference standards and three samples of material from Reference 1.
The data presented a challenge because Reference 2 was actually a mixture of Reference 1 and 3. Consequently, one would expect the samples to appear more similar to Reference 1 than 3, but there may be some similarity of the samples to Reference 2. The data from each sample and reference material was analyzed by the methodology described above. Sample statistics were collected on the distribution of differences between the reference material and each corresponding sample and these are presented in Table 1. These are the summary statistics from the distribution D(S) calculated for each pair that consists of a sample and a reference with from the d-lines weighted by intensity. Also, histograms were then generated to visually assess the distribution of differences between the sample and reference material, and a selected example is shown in Figure 1.

From Table 1 it can be seen that for each sample the variance, mean, and median of the distribution D(S) were smaller when the sample was paired with Reference 1 than with Reference 3. This is the correct conclusion. The minimum of the distribution did not follow this conclusion, but the minimum is computed from a single subset so that one would expect its results to be somewhat unpredictable. Our most favored distribution summary is the mean. Two of the three samples showed a smaller mean when paired with Reference 1 than with Reference 2, but the third sample had a slightly lower mean when paired with Reference 2. Still, these means were much lower than when a sample was paired with Reference 3.

Figure 1 shows a histogram of the distribution of D(S) for one sample paired with Reference 1 and Reference 3. It shows that D(S) is not a symmetric distribution, but that the fit to Reference 1 generates a distribution closer to zero than the fit to Reference 3. Because of the lack of symmetry it is expected that a better summary measure can be selected to match the sample to a reference than the more traditional measures (mean, median, variance) studied here. Further work to understand, refine, and improve these measures continues.

Table 1: Summary statistics on the distribution of differences between Sample & Reference

Variance			
	Sample 1	**Sample 2**	**Sample 3**
Reference 1	3.797423e+12	5.362168e+12	7.395444e+12
Reference 2	3.810584e+12	13.36692e+12	2.028746e+12
Reference 3	22.17329e+12	144.0470e+12	230.4793e+12
Minimum			
	Sample 1	**Sample 2**	**Sample 3**
Reference 1	5034	5305	242
Reference 2	1339	2285	68
Reference 3	759	52	49
Median			
	Sample 1	**Sample 2**	**Sample 3**
Reference 1	456,300	495,800	56,600
Reference 2	415,100	177,800	11,820
Reference 3	691,100	546,400	330,900
Mean			
	Sample 1	**Sample 2**	**Sample 3**
Reference 1	844,700	1,729,000	1,379,000
Reference 2	1,021,000	2,683,000	507,300
Reference 3	2,507,000	4,175,000	4,308,000



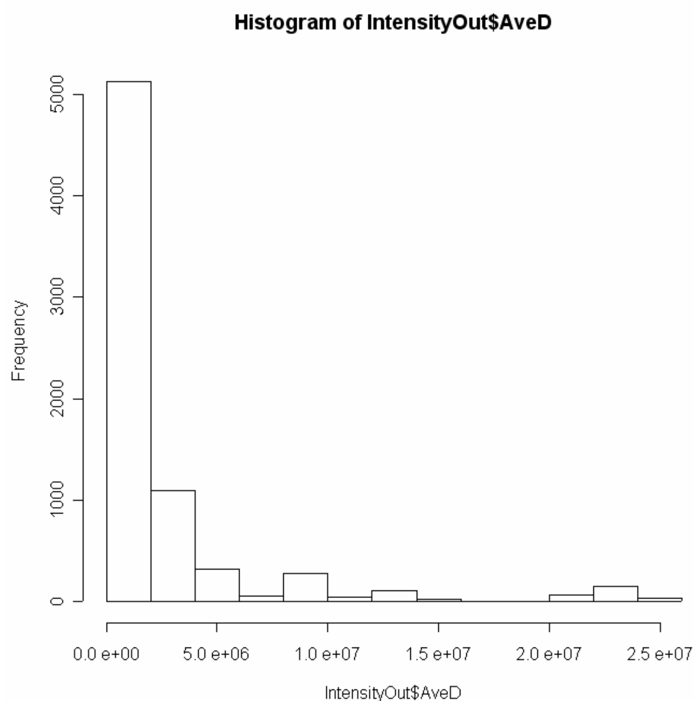


Figure 1: Sample 2 with Reference 1 and Reference 3

5. Conclusion

A new methodology was developed to compare XRPD samples and references. A metric was created and a d-line distribution of differences was used to establish whether a true difference exist between a reference and sample material. This approach is unique in that it makes use of a novel distance metric along with a distribution generated from a computational methodology that would not have been possible prior to the advances in computer technology. Whereas, current approaches focus on grouping or clustering to match a sample to a reference. It also focuses on the d-lines instead of creating direct models for intensity. Several sample chemical material d-lines were obtained and compared to a published reference source to determine if a definitive difference exists.

Each of the samples was distinguished from the three references accurately. In this particular series of experiments the results emphasized the need for encompassing intensity into the metric, but only in the form of weights. Further research is needed in this regard to determine alternate weighting schemes to enhance the final d-line distribution as well as alternative metrics to compare the skewed distributions that result.

The purpose of this paper was to outline the methodology to measure differences between samples, via the score or metric of d-line differences between samples. This methodology provides a scientific answer to the legal question, is the material being tested the same to what is patented? However, the methodology does not address the tolerance in case peaks are shifted. A confidence window for how far from a given d-line one would accept within a sample needs to be established. This window must reflect the various sources of variability in the process. One-way to establish a quantitative rule-of-thumb for a confidence window is to use a Gage Repeatability and Reproducibility (R&R) design. This is a non-biased empirical approach that accounts for the various sources of variation in batches, operators, equipment and its associated geometry (e.g., Bragg-Brentano, Debye-Scherrer, etc.). Of which the later is often referenced as a problematic noise contributor for measured peak position [8]. The use of Gage R&R is more representative than the use of first principles to handle tolerance, in that the latter is likely to result in a biased estimate of what one should expect the error function to look like. We feel that the importance of refining the weights and determining an appropriate confidence window will result in a definitive

methodology for establishing equivalence between sample material and a reference. The use of the methodology could easily be automated and could likely grow as its use becomes more prevalent in the pharmaceutical community.

References

1. T.G. Fawcett, J. Faber, C.R. Hubbard, "*Formulation Analyses of Off the-Shelf Pharmaceuticals*"s, American Pharmaceutical Review, 2005, 80-82 118.
2. J. Faber, C.A. Weth, J. Bridge, "*A Plug-In Program to Perform Hanawalt or Fink Search-Indexing Using Organics Entries in the ICDD PDF-4/Organics 2003 Database*, Powder Diffraction, **19** (1), March 2004.
3. S. Agatonovic-Kustrin, V. Wu, T. Rades, D. Saville, I.G. Tucker, *Ranitidine Husdrochloride X-Ray Assay Using a Neural Network*, Journal of Pharmaceutical and Biomedical Analysis, **22**, (2000), 985-992.
4. G. Berti, "*Modeling and Optimization Algorithm to Analyse XRPD Data via Modulation and Pseudo-Voigt Functions*, Advances in X-Ray Analysis, **39,** (1997), 465-471.
5. G.C. Runger, F.B. Alt "Choosing Principal Components for Multivariate Statistical Process Control", *Communications in Statistics—Theory and Methods* , 1996, 25(5), 909-922.
6. K. Jorgensen, V. Segtnan, K. Thyholt, T. Naes, "*A Comparison of Methods for Analyzing Regression Models with Both Spectral and Designed Variables*, Journal of Chemometrics, 2004, **18**, 451-464.
7. A. Antoniadis, J. Berruyer, A. Filhol, "*Maximum-Likelihood Methods in Powder Diffraction Refinements*, Acta Cyrstallographica, 1990, **A46-8**, 692-711.
8. S. R. Byrn, S. Bates, "*Regulatory Aspects of X-Ray Powder Diffraction, Part 1*, American Pharmaceutical Review, 2005, 55-59.