A PLUG-IN PROGRAM TO PERFORM HANAWALT OR FINK SEARCH-INDEXING USING ORGANICS ENTRIES IN THE ICDD PDF-4/ORGANICS 2003 DATABASE

J. Faber, C. A. Weth and J. Bridge*

International Centre for Diffraction Data (ICDD)
Newtown Square, PA 19073, USA
**West Chester University, Department of Computer Science,*
West Chester, PA 19380

ABSTRACT

In an attempt to fill a gap between fully automatic search/match programs and purely manual methods based on paper products, a relational database plug-in has been developed that functions as a PC-based Search/Index program for extracting information from PDF-4 powder diffraction databases. The plug-in provides an adjustable search window and match window to account for experimental errors. Both Hanawalt [1,2] and Fink [3] search methods are incorporated. In this paper, we report search-indexing results obtained with the new PDF-4 plug-in applied to a new relational database, the PDF-4/Organics 2003. This database has 24,385 experimental entries and 122,816 calculated patterns derived from the Cambridge Crystallographic Database (CSD). We introduce a Goodness of Match (GOM) parameter to describe the relative agreement between the experimental input data and selected reference patterns from the PDF-4/Organics 2003. The relevance of the GOM is illustrated in several example problems. Multiphase samples can be treated on a phase-by-phase basis.

INTRODUCTION

The International Centre for Diffraction Data (ICDD) has been the primary reference for X-ray Powder Diffraction (XRPD) data for over 50 years. The primary information in the PDF is the collection of d-I data pairs, where the d-spacing (d) is determined from the Bragg angle of diffraction, and the peak intensity (I) is obtained experimentally under the best possible conditions for a phase-pure material. These data provide a data mining [4-5] capability as well as "fingerprint" of the compound because the d-spacings are fixed by the geometry of the crystal and the intensities are dependent on the contents of the unit cell. Hence, d-I data may be used for identification of unknown materials by locating matching d-I data in the PDF with the d-I pairs obtained from the unknown specimen. Identification is the most common use of the PDF, but the presence of considerable supporting information for each entry in the PDF allows further characterization of the specimen. Examination of the crystal data, Miller indices, intensity values, scale factors, physical property data and the comprehensive literature reference data provide extraordinarily useful information concerning the specimen under study. For pharmaceutical R&D, XRPD and the PDF have been used for example as an indispensable tool in phase identification (both qualitative and quantitative), in the identification of unknowns, evolution of polymorphism and solvate structures, and crystallinity determinations. The impact of the PDF as a reference pattern database has been used in patent disclosures and as such has immediate impact for pharmaceutical R&D.

The PDF has exhibited recent dramatic growth in entry population over the past 5 years. Historically, the PDF-2 has been a flat file database that contains powder patterns of inorganic compounds. However, in late 1998, the ICDD reached an agreement with the Cambridge Crystallographic Data Center (CCDC) that allows for the calculation of x-ray powder patterns from the structural information in the Cambridge Structural Database (CSD). The resultant explosion in the organic population from 25,000 to 150,000 entries in 2003 is a direct result of this agreement. A completely new relation database (RDB) was used to house the new PDF-4/Organics 2003. The principal classes of database compounds are organic and organo-metallic. The properties of new PDF-4 databases are illustrated in Table 1.

We could anticipate that some search-indexing problems may arise using the PDF-4/Organics database:

- Only organic entries are present in the PDF-4/Organics 2003 database. However, note that 1,117 inorganic compounds are present in the PDF-4/Organics 2004. Inorganic excipients are particularly relevant for pharmaceutical R&D.
- There could be larger uncertainties in the lattice parameters since single crystal experiments are often not optimized for high-resolution d-spacing determinations. The focus is on integrated intensities. Only 623 calculated pattern entries from the CSD indicate that cell constants were obtained using powder diffraction methods.
- Organic entries are often done at low temperature. Comparison between low-temperature reference data and room temperature powder diffraction data does not account for thermal expansion in the reference pattern.
- Organic powder diffraction patterns often contain substantial preferred orientation effects. However, as we shall see, we have implemented rotation operators that permute the strongest lines in the pattern (in both Hanawalt and Fink analyses), which has the effect of taking preferred orientation into account. Severe preferred orientation effects cannot be overcome since this would completely distort these strong line/long line methods. We will discuss this issue in more detail.

The focus of this paper is to present applications that demonstrate the power of a PDF-4/Organics 2003. We shall demonstrate this analytic power by illustrating results obtained from phase identification and search-indexing, using Hanawalt and Fink methods. A preliminary report for PDF-4/Full File 2002 (a predominantly inorganic database) has been given [6].

	PDF-4/ Full File 2003	PDF-4/ Organics 2003	PDF-4/* Organics 2004
Organic Compounds	**25,609**	**147,201**	**217,077**
Inorganic Compounds	**133,370**		**3,048**
Both Organic and Inorganic	**1,931**	**1,776**	**1,931**
Only Inorganic	**131,439**		**1,117**
Calculated patterns from CSD		**122,816**	**191,468**
Drug Activity Index		**4,508**	**6,343**
Pharmaceuticals	**2,039**	**1,192**	**2,039**
Excipients	**801**	**184**	**1114**
Forensic Materials	**3,767**	**2,015**	**2,113**
Pigments	**342**	**284**	**296**
I/Ic	**73,087**	**125,342**	**195,316**
Total Entries	**157,048**	**147,201**	**218,194**

Table 1. Selected entry counts of the PDF-4 databases. Please note that PDF-4/Organics 2004 will be released in November, 2003. Please note that because entries can be listed in both the inorganic and organic collection, the total number of distinct entries is obtained from the organic and only inorganic rows in the Table.

PDF-4/ORGANICS 2003

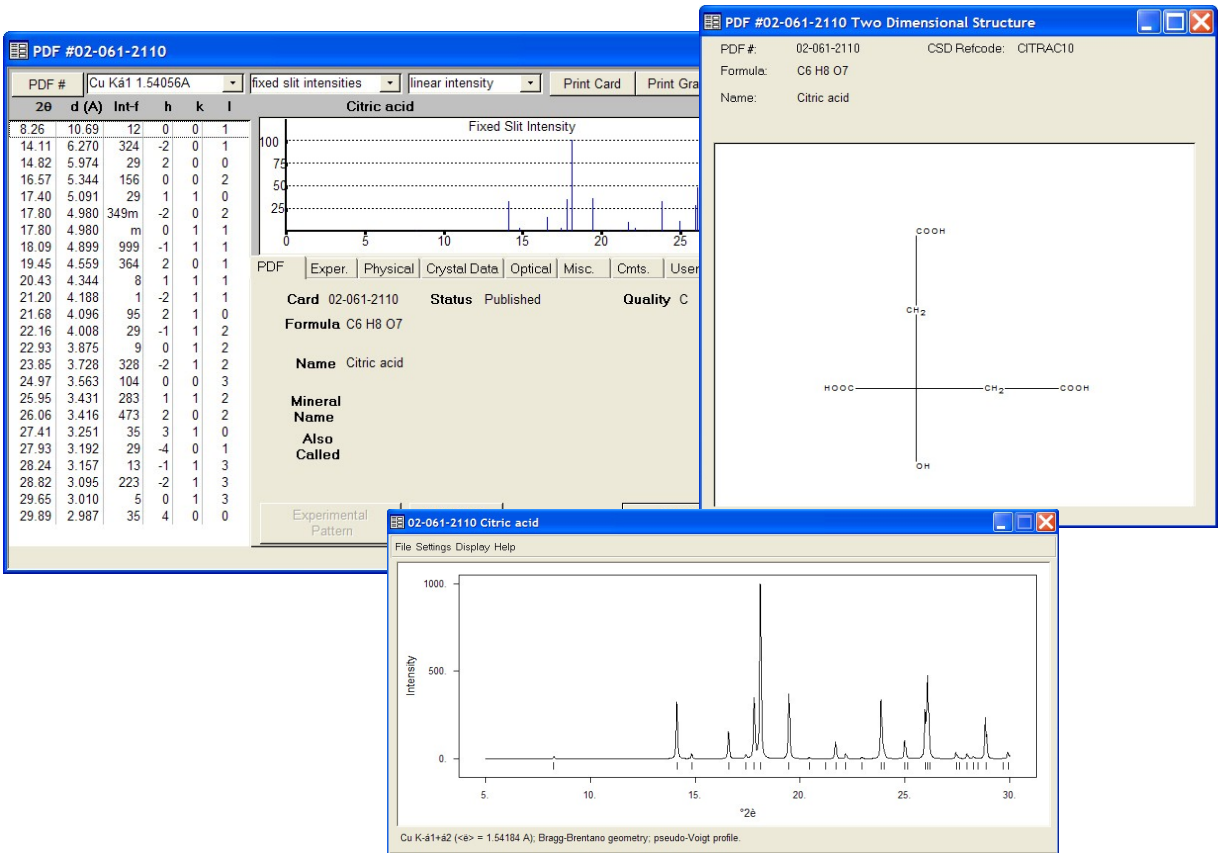

Figure 1. Example data from the PDF-4/Organics 2003 for Citric Acid. Note the 2D structure display and the on-the-fly digitized pattern.

The PDF-4 database contains interplanar spacings (d) and relative intensities (I). However, other useful data such as synthesis, physical properties and crystallographic data are also stored in the database. With this new format, we will provide a broader range of analyses, for example, improved quantitative analyses, full pattern display, bibliographic cross referencing, etc. The PDF-4 uses relational database technology that provides pliable access to the database to carry out data mining studies and enhances the pursuit of conventional materials characterization using diffraction techniques (see Faber et al. [5]). In addition to better access to some of the RDB fields, users can also build search criteria by combining individual search conditions using Boolean operators. The availability of logical operators for combining the search condition is very useful in arriving at the desired information from the database

The CSD database is being used to calculate entries in the PDF. Thus, to derive d-spacing and peak intensity data requires the synthesis of full diffraction patterns, i.e., we use the structural data in the CSD database and then add instrumental resolution information. In addition to the peak intensities, $|F(hkl)|^2$, the square of the structure factor magnitudes will also be calculated. Thus, calculated powder patterns are obtained for all CSD entries in the PDF-4/Organics 2003 RDB. For example, we can calculate (on-the-fly) a selected profile function to describe paracrystallinity, or particle size and/or strain effects. PDF data for an ideally random crystal distribution in the absence of preferred orientation may also be obtained. In the future, preferred orientation models will be developed. The main focus is to provide tools that can be used for materials design.

The CSD contains bitmap control integers that can be used to project out specific categories of entries in the CSD. Of particular interest for pharmaceuticals is the drug activity flag. There are approximately 250,000 entries in the CSD and of these, approximately 8000 have the drug activity flag set. The PDF-4/Organics 2003 contains calculated patterns for 4, 292 of these entries. The process of calculating PDF data is an ongoing task; we will calculate powder patterns for all entries in the CSD when the ICDD editorial review has been successfully completed.

SEARCH-INDEXING USING THE PDF-4/ORGANICS RDB: HANAWALT AND FINK SEARCH/MATCH PROCEDURES

Most of the commercial software packages for qualitative phase identification have been designed to implement fully automatic search/match sequences [6-17]. On the other hand, traditional methods of search/match (based on *d*-spacings, intensities and chemistry) are mainly manual techniques using paper-based search/indices. Manual techniques were first discussed by Hanawalt and these persist for a variety of reasons. The search-indexing plug-in discussed here follows a traditional path to act as a replacement for paper search manuals published by the ICDD. An advantage to this approach is that Hanawalt and Fink methods can be followed in great detail as search-indexing proceeds. The educational benefit of this approach is also realized.

Traditional methods for search/match in powder diffraction are based upon combinations of d-spacing, intensities and chemistry. Table 2 lists the types of search indices currently being used.

Table 2. Types of Data Search Indexes

Index	**Entry Method**	**Search Parameters**
Alphabetic	Chemistry, chem. formula fragments	Permuted chemical formula fragments
Hanawalt	d,I pairs, sorted in decreasing intensity	3 strongest lines
Fink	d,I pairs, sorted in decreasing d-space	8 longest lines (longest of the strongest)

The Hanawalt search method has been implemented for many years at the ICDD. The method involves sorting the patterns in the PDF according to the d-spacing value of the strongest line. This list is broken into discrete d-space intervals defined as Hanawalt groups. A small overlap in d-intervals is employed to reduce the probability of missing powder pattern entries due to uncertainty in the d-space accuracy. Each Hanawalt group is sorted in order of decreasing d-spacing of the second most intense diffraction line. Subsequent lines are listed in order of decreasing intensity. The analysis rests on the three most intense lines, but the eight most intense lines are listed. Considerable redundancy exists in this method because patterns appear twice for the (1,2) and (2,1) pairs when $I_2/I_1 > 0.75$ and $I_3/I_1 > 0.75$. Patterns appear three times (1,2), (2,1), and (3,1) when $I_3/I_1 > 0.75$ and $I_4/I_1 < 0.75$. The rationale for multiple entries is to minimize problems of preferred orientation, especially when these affect the three strongest lines. In summary, the Hanawalt method relies on the d-spaces for the three strongest lines; further confirmation of a search hit is taken from matches on the eight strongest lines.

The Fink method was designed as an index based on the eight strongest d-spaces in the experimental pattern, but these are ordered in decreasing d-spacing. In short, the Fink method considers the 8 longest of the strongest diffraction lines. Creating permutations of these is not practicable for large databases as the corresponding paper manuals become enormous. However, as we shall see, permuting the Hanawalt three strongest lines or the Fink eight longest lines is straightforward using computer methods. For both the Hanawalt and Fink methods, the problem is that the associated paper manuals have grown cumbersome and difficult to use. In addition, the integration of elemental composition and other important ancillary information is not easily accomplished with these methods. Filtering criteria need to be "remembered" while carrying out the search/indexing process. We have developed a "plug-in" for the PDF-4 databases that implements the Hanawalt/Fink strategy, including chemistry, subfile and quality-mark filters.

SEARCH-INDEX PLUG-IN

The basic idea of the plug-in is to provide d,I pairs as input to the program. The d-spaces are in Å and the I's are peak intensity values from the x-ray powder diffraction experiment. As additional input, P is a phase parameter associated with each d,I pair; if P=1, the peak is included in the analyses, otherwise the peak is ignored. As we shall see, this is quite helpful in multiphase problems. Also, contaminant peaks can be easily excluded in the analysis by adjusting P. The principal input is from an ASCII file that contains the d,I pairs. However, the plug-in can also accommodate 2θ,I pairs if the first ASCII record also contains the wavelength. The uncertainty in d, Δd, can be obtained by taking the derivative of Bragg's Law:

$$\Delta d = d \cdot \cot\theta \cdot \Delta\theta . \qquad \text{Eq. 1}$$

In the case of the Hanawalt method, the search window, $SW = \Delta(2\theta_{SW})$ defines the Hanawalt group. The angular dependence of the search window and match windows are defined by Eq. 1. The match window, $MW = \Delta(2\theta_{MW})$ defines the PDF entry lines that match with the experimental data. Up to eight strongest experimental lines appear as d1 – d8 just above the match list box in Figure 2. The best match is assigned to PDF # 000161157. For this match, at least 4 of the 8 strong lines are matched as discussed below.

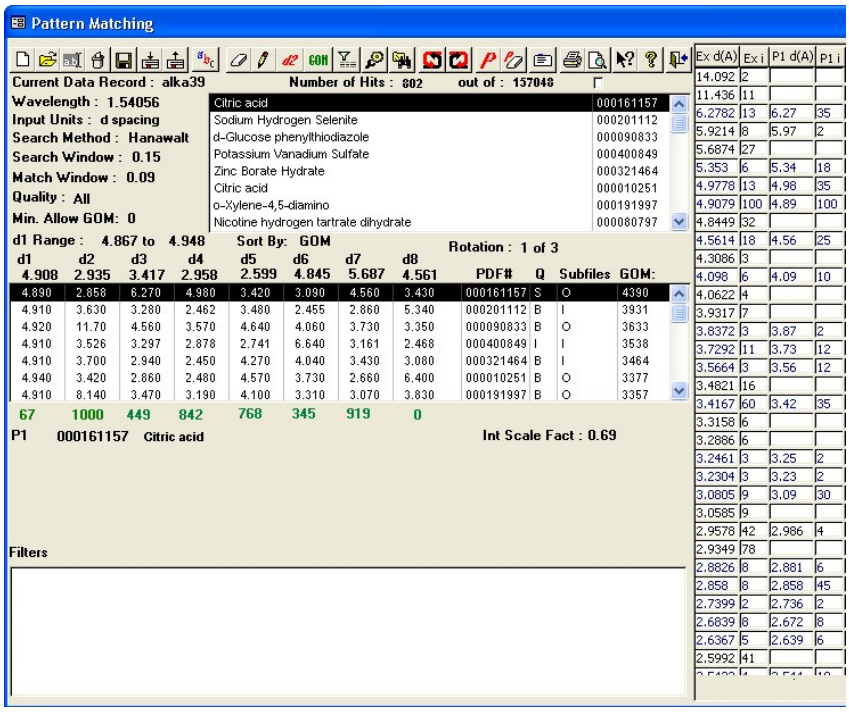

Figure 2. Hanawalt method applied to an over-the-counter medication. The drug is Alka-Seltzer Plus, normally ingested after dissolution in water. The tablets were ground and standard XRD experiments were performed. A peak-listing program was used to define d-spaces and peak intensities for all Bragg lines detected.

Match window hits are selected and an algorithm is used to obtain the hit that best matches the experimental data. This best match is obtained by calculating the GOM, defined by

$$GOM = 1000 \cdot \sum [1 - |\delta d| / SW]^2 , \qquad \text{Eq. 2}$$

where δd is taken from the difference between the d-spacing for the unknown and the d-spacing for the candidate reference pattern, the sum is taken over the experimental lines and their corresponding match lines in the selected PDF entry, and SW is the search window defined from Eq. 1. Notice that a perfect match between experiment and the PDF for 8 lines would yield GOM=8000. Also, GOM values <1000 are not significant since this corresponds to the identification of only a Hanawalt Group. GOM <2000 means that no single reasonable entry has been identified within the Hanawalt Group. For the analysis presented in Figure 1, GOM=4390.

The Intensity Scale Factor seen in Figure 2 is calculated based on all matched lines in the analysis.

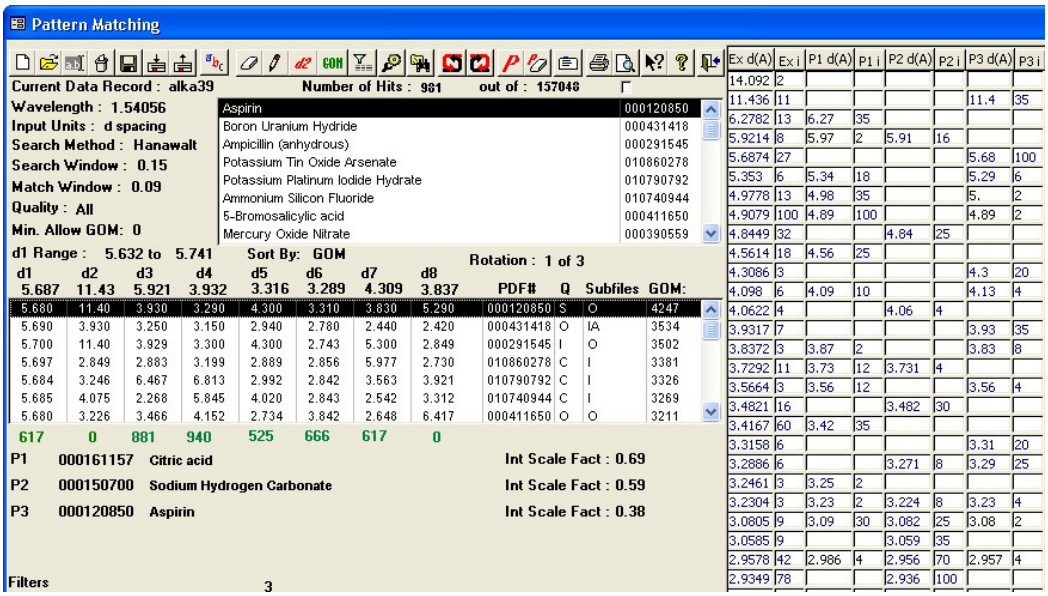

Figure 3. Two additional phases were identified in this analysis: a second and third search were performed. The GOM=5140 for the second phase (not shown) and GOM=4247 for the third phase.

The GOM=4247 (Figure 3) indicates that nearly 4 of the 8 strongest lines have been matched. In this case, all apart from one weak line (d=14.092, I=2) have been account for. Notice also in Figure 2 that relative scaling factors are indicated for each detected phase. However, these are only approximate since RIRs (reference intensity ratios) are not used in this analysis. Obtaining semi-quantitative analyses using reference intensity ratios is under development. One of the powerful features of the plug-in is that the data grid on the right side of Figs. 2-3 can be filled by any selection from the hit list. The data grid lists all experimental lines and all lines from the selected entry in the hit list. Thus, a detailed comparison between the match hit and the input data can be seen at a glance.

Enhanced search indexing is realized by allowing for permutation of the order of the experimental lines, thus removing the need for redundant (and complex) indexing of reference data from the database. In the Hanawalt search method we have implemented rotation operators to permute the order of the three strongest lines. This is indicated by the rotation counter: "Rotation: 1 of 3", seen in Figs. 2 and 3. The rotation operator is also available for the Fink method (not shown), however, in this case since we are ordering 8 strongest of the longest d-space lines, the range is 1-8. The utility of these rotation operators is that they alter the order of the strongest lines and hence allow moderate degrees of preferred orientation to be taken into account. It has been our experience that when intensities are problematic, the Fink method tends to help resolve the component phases. We have implemented a third method, designated "long 8" that completely ignores the peak intensities. The utility of this third method in currently under evaluation.

SUMMARY

We have illustrated several examples of the complementary use of XRPD techniques coupled with a new organic database, the PDF-4/Organics 2003 RDB. An example, Citric Acid, was used to show some of the power features available in this new PDF-4/Organics. In particular, calculated on-the-fly powder patterns were generated and 2D structures are available. The search-indexing example, Alka-Seltzer Plus, was used to show search-indexing results using Hanawalt and Fink methods for phase ID. In this case, we were able to identify the three most abundant components in the tablet. We feel that the importance of the PDF-4/Organics 2003 RDB will grow as its use becomes commonplace in the pharmaceutical community.

REFERENCES

[1] Hanawalt, J. D. and Rinn, H. W., Ind. Eng. Chem. Anal. 8, 244 (1936); Hanawalt, J. D., Advances in X-Ray Analysis 20, 63-73 (1976).
[2] Hanawalt, J. D., Cryst. in North America, Apparatus and Methods, American Crystallographic Association, Chapter 2, 1983, pp.215-219.
[3] Bigelow, W. and Smith, J.V., ASTM Spec Tech Publ. STP 372, 54-89 (1965).
[4] Faber, J., Kabekkodu, S.N. & Jenkins, R. (2001), *International Conference on Materials for Advanced Technologies*, Singapore, unpublished; Kabekkodu, S.N., Faber, J. and Fawcett, T., Acta Cryst., Vol. B58, 333-337 (2002).
[5] Faber, J. and Fawcett, T., Acta Cryst., Vol. B58, 325-332 (2002).
[6] Faber, J., Weth, C. A. and Jenkins, R. (2001), Materials Science Forum Vol. 378-381, 106-111 (2001).
[7] Johnson, G. G., Jr., and Vand, V., Ind. Eng. Chem., Vol. 59, 19 (1965).
[8] Nichols, M. C., Lawrence Livermore Lab. Report UCRL-70078 (1966).
[9] Frevel, L. K., Adams, C. E. and Ruhberg, L. R., J. Appl. Crystallogr., Vol. 9, 300-305 (1976).
[10] Marquardt, R. G., J. Appl. Crystallogr., Vol. 12, 629-634 (1979).
[11] Snyder, R. L., Advances in X-Ray Analysis Vol. 24, 83-90 (1980).
[12] Jobst, B. A., Goebel, H. E., ibid 25, 273-282 (1981).
[13] Parrish, W., Ayers, G. L., Huang, T. C., ibid Vol. 25, 221-229 (1981).
[14] Jenkins, R., Hahm, Y., Pearlman, S. and Schreiner, W. N., ibid Vol. 23, 279-285 (1979).
[15] Goehner, R. P. and Garbauskas, M. F., X-Ray Spectrom. Vol. 13, 172-179 (1984).
[16] Toby, B. H., Powder Diffraction Vol. 5, 2-7 (1990).
[17] Caussin, P., Nusinovici J. and Beard, D. W., Advances in X-ray Analysis, Vol. 31, 423-430 (1987); ibid Vol. 32, 531-538 (1988); Nusinovici J. and Winter, M. J., ibid Vol. 37, 59-66 (1993).