Organic Process Research & Development **2002**, *6*, 317−322

Technology Reports

Solution Concentration Prediction for Pharmaceutical Crystallization Processes Using Robust Chemometrics and ATR FTIR Spectroscopy

Timokleia Togkalidou,[†] Hsien-Hsin Tung,*[,‡] Yongkui Sun,[‡] Arthur Andrews,[‡] and Richard D. Braatz[†]

University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Box C-3, Urbana, Illinois 61801, U.S.A. and Merck and Co., Inc., 126 East Lincoln Avenue, P.O. Box 2000, Rahway, NJ, 07065, U.S.A.

Abstract:

In the pharmaceutical industry, a vast number of compounds are produced by solution crystallization, making the design and development of such processes of critical importance. The kinetics of crystal growth and nucleation, the fundamental mechanisms of a solution crystallization process, are strongly dependent on supersaturation (the difference between solution concentration and the saturation concentration). The present study uses attenuated total reflection (ATR) Fourier transform infrared (FTIR) spectroscopy, coupled with robust chemometric techniques, for the on-line measurement of solution concentration of two pharmaceutical compounds in multicomponent systems in the presence of impurities and over a wide range of temperature. To our best knowledge, this is the first time that ATR FTIR spectroscopy has been applied to a multicomponent pharmaceutical system. The resulting models show high accuracy in predicting the solution concentration and are applied successfully in measuring the solubility for the cases of cooling and antisolvent crystallization.

Introduction

In the crystallization process, the measurement and control of supersaturation, which is the difference between the solution concentration and saturation solution concentration, are critical. Important parameters in pharmaceutical industry such as crystal size distribution (CSD) and crystal shape of the product pharmaceuticals are determined by the supersaturation profile followed during crystallization. Unfortunately, it is challenging to accurately measure the solution concentration in dense crystal slurry typical of industrial operations. Periodical off-line sampling is tedious in the laboratory and is subject to various disturbances in a production environment. In situ infrared (IR) measurement avoids the problem of sampling and can be applied in multiple component systems. The feasibility for the in situ measurement of solution concentration in dense crystal slurries has been demonstrated in various studies in the literature.[1−3] However, the accuracy of this method can drop drastically when the IR spectra are sensitive to temperature and the presence of other compounds in the solution.[4] To address these issues, inferential methods are available which apply chemometric techniques over selected ranges of IR spectra.[5−7] More recently it was demonstrated that attenuated total reflection (ATR) Fourier transform infrared (FTIR) spectroscopy coupled with robust chemometrics gives measurements of solution concentrations with high accuracy in a potassium dihydrogen phosphate−water slurry, for a wide range of temperature.[8] In this report, this methodology is applied to two drugs: compound-X and compound-Y (the names and structures of the compounds cannot be released for propriety reasons). This appears to be the first time that accurate solution concentration measurements have been reported for a pharmaceutical crystallization process over a complex array of parameters including temperature, solvent composition, and raw material impurities. It is the first time that robust chemometric techniques coupled with ATR FTIR spectroscopy is applied to a multicomponent pharmaceutical system for the prediction of solution concentration and the determination of the solubility curve. It would be very interesting to apply this technique not only for solubility curve determination but particularly for supersaturation control.

Theory

Principal component regression (PCR) and partial least-squares are applied to a set of input to output variables. The input or predictor variables are the absorbances at distinct wavenumbers and the temperature, and the output or

* To whom correspondence should be addressed: Merck & Co., Inc., 126 East Lincoln Avenue, P.O. Box 2000, Rahway, NJ, 07065. Telephone: (732) 594-6686. Fax: (732) 594-9140. E-mail: hsien_hsin_tung@merck.com.

[†] University of Illinois at Urbana-Champaign.

[‡] Merck & Co., Inc.

(1) Dunuwila, D. D.; Caroll, L. B., II.; Berglund, K. A. *J. Cryst. Growth* **1994**, *137*, 561−568.
(2) Dunuwila, D. D.; Berglund, K. A. *J. Cryst. Growth* **1997**, *179*, 185−193.
(3) Lewiner, F.; Fevotte, G.; Klein, J. P.; Pfefer, G. *Proceedings of the 14th International Symposium on Industrial Crystallization*; Institution of Chemical Engineers: Rugby, U.K., 1999; Paper 47.
(4) Bellon-Maurel, V.; Vallat, C.; Goffinet, D. *Appl. Spectrosc*. **1995**, *49* (5), 563−568.
(5) Workman, J. J.; Mobley, P. R.; Kowalski, B. R.; Bro, R. *Appl. Spectrosc. Rev.* **1996**, *31*, 73−124.
(6) Mobley, P. R.; Kowalski, B. R.; Workman, J. J.; Brov, R. *Appl. Spectrosc. Rev.* **1996**, *31*, 347−368.
(7) Bro, R.; Workman, J. J.; Mobley, P. R.; Kowalski, B. R. *Appl. Spectrosc. Rev.* **1997**, *32*, 237−261.
(8) Togkalidou, T.; Fujiwara, M.; Patel, S.; Braatz, R. D. *J. Cryst. Growth* **2001**, *231*, 534−543.

10.1021/op015516x CCC: $22.00
Published on Web 04/27/2002

predicted variable is the concentration (or solvent composition). The inclusion in the model of multiple absorbances, instead of distinct peaks, averages the measurement noise and allows for the explicit consideration of peak shifts. Ordinary least-squares (OLS) cannot be applied due to a number of factors. First, it may occur that the number of unknowns is greater than the number of experiments. Additionally, the input variables may be correlated to each other, introducing singularities in the OLS method and consequently high levels of inaccuracy. Furthermore, variables that do not carry any information but are correlated with noise would be falsely included in an OLS model. In PCR, the set of input variables is transformed to a new set of orthogonal uncorrelated input variables called principal components, a subset of which is used to construct a regression model to the predicted variable. The principal components associated with noise are discarded. There are several variants on PCR,[9] which use different procedures to select which of the principal components to include in the regression model. Different PCR variants as well as PLS can perform better or worse, depending on the particular data set. In the robust chemometric approach,[8] a number of chemometric methods are applied, and the most accurate model produced by these methods is selected. The chemometric methods include top-down selection PCR (TPCR),[9] correlation PCR (CPCR),[9] forward selection PCR 1 (FPCR1),[10] forward selection PCR2 (FPCR2),[9] confidence interval PCR (CIPCR),[11] and partial least-squares (PLS).[12]

The experimental data are subdivided into a calibration and validation set in six different ways. Half of the data points are used as the calibration set to construct the model. The other half of the data points are used to evaluate the model. The resulting 36 combinations (6 subdivisions by 6 methods = 36) are evaluated by using the width of the prediction interval[13] and the relative ability of prediction. The subdivisions are used to help locate any biases in the data. As a rule, accurate models produce small prediction intervals (a small error in the prediction) and a relative ability of prediction close to unity.

Experimental Setup

In-situ measurements of solution concentration and solubility were performed using spectra obtained by a attenuated total reflectance (ATR FTIR) immersion probe attached to an ASI Applied Systems ReactIR 1000 FTIR spectrophotometer connected to a Pentium computer, running ReactIR Software version 2.

Samples were prepared using appropriate amounts of compound-X, toluene, and *n*-heptane for the first system, and compound-Y, ethyl acetate, 2-propanol, *n*-heptane, and water for the second system. The samples were stirred using an overhead Cole−Palmer Instrument mixer. The sample temperature was controlled using a HAAKE C50 water/glycol bath and was measured using a temperature monitoring system attached to the FTIR spectrophotometer, the sensor being a 100Ω Platinum LTD, stainless steel with Teflon coating. The FTIR spectra were collected in the range 650−4000 cm^{-1}. The water content was measured using a Brinkmann 684 KF-coulometer.

In the following section of this report, due to the diverse scenarios encountered by these two compounds, the experimental procedure and results of each scenario are put together to facilitate the flow of information.

Table 1. **Number of spectra, temperature, and concentration ranges for five solutions**

	solution concentration (g/kg solution)	temperature range (°C)	number of spectra
1	33.4	20−56	25
2	76.8	30−64	25
3	115.2	38−65	25
4	152.3	40−75	25
5	185.1	50−70	25

Procedure and Results

1. Compound-X in Single-Solvent (Toluene) System with Varying Temperature. Compound-X is polymorphic, existing in several different crystal forms in the solid phase. Each of the forms has a distinct solubility curve. The most thermodynamically stable form at the operating conditions considered here is form V. To build the inferential (calibration) model, five solutions of known concentration of form V were prepared. Spectra were collected for a broad range of temperature. Table 1 reports the number of spectra, the corresponding solution concentration range (g/kg of solution), and the temperature range for each of the five solutions. The total number of spectra that were collected was 125. Pure toluene at 25 °C was used as the background spectrum.

The above 125 spectra were used to construct the calibration curve, that is, the inferential model. The wavenumber range selected for the calibration was 798−1458 cm^{-1}, which are wavenumbers in which compound-X absorbs strongly. The chemometric model produced by the CPCR method gave the smallest prediction interval. The model predicted compound-X concentration with an accuracy of ±1.2% or ±1 g/kg solution from 20 to 60 °C, at the 95% confidence level.

After the inferential model was constructed, solubility experiments were conducted to determine the solubility from 20 to 54 °C. A slurry of compound-X in toluene at 20 °C was created, and spectra were collected once the spectra reached a plateau. The temperature was increased from 20 to 25 °C in 70 min, and the mixture was left to equilibrate for 30 min. Spectra were collected every 5 min. The temperature then was raised to 34 °C, and the procedure was repeated. Measurements of IR spectra were taken at 20, 25, 34, 43, and 53 °C. For each temperature, spectra were collected for at least 30 min every 5 min. In each case the spectra reached a plateau before increasing the temperature further. The temperature profile is shown in Figure 1. Also, samples were taken from the slurry and tested with X-ray

(9) Xie, Y.; Kalivas, J. *Anal. Chim. Acta* **1997**, *348*, 19−27.
(10) Draper, N. R.; Smith, H. *Applied Regression Analysis*, 2nd ed.; John Wiley & Sons: New York, 1981.
(11) Togkalidou, T.; Braatz, R. D. *Proceedings of the American Control Conference*; IEEE Press: Piscatawa, New Jersey, 1999; pp 2548−2552.
(12) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: New York, 1991.
(13) Faber, K.; Kowalski, B. R. *Chemom. Intell. Lab. Sys.* **1996**, *34*, 283−292.

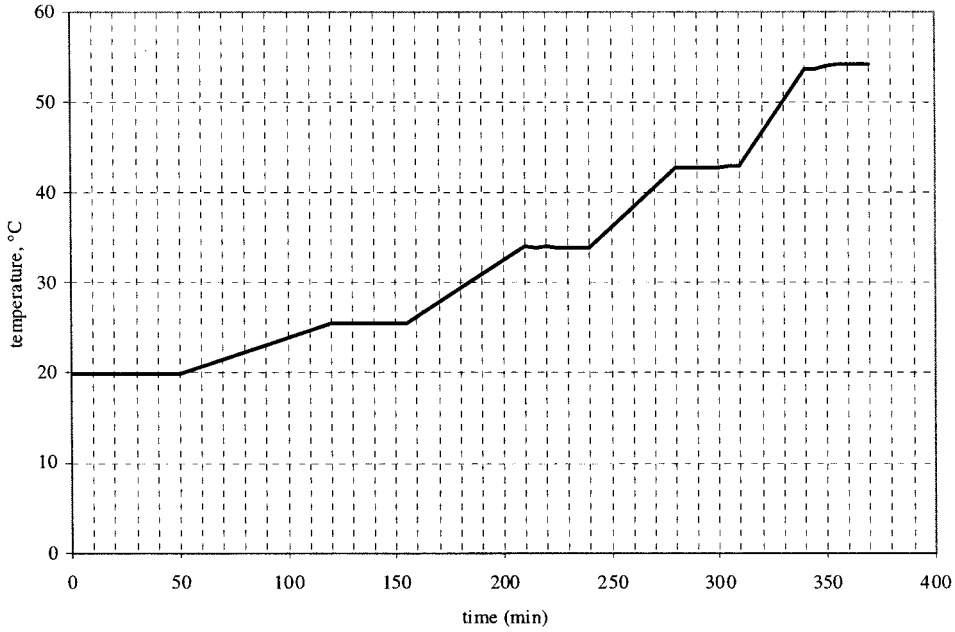

Figure 1. **Temperature profile applied to the system for the solubility curve determination.**

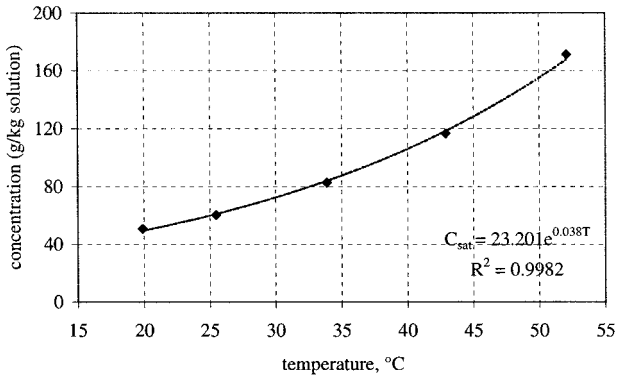

Figure 2. **Solubility of compound-X in toluene as a function of temperature.**

Table 2. **Number of spectra, and concentration ranges for 10 solutions**

	solution concentration range (g/kg solution)	solvent/total volume	temperature range (°C)	number of spectra
1	26.25−43.08	0.55−1.00	18.9−20.1	6
2	2.56−5.23	0.43−1.00	19.7−20.7	9
3	17.57−24.30	0.67−1.00	19.5−20.6	4
4	9.08−12.80	0.56−0.71	20.1−20.9	5
5	26.24−33.58	0.75−1.00	20.0−20.2	3
6	13.34−15.69	0.65−0.77	19.2−20.3	4
7	10.75−11.20	0.63−0.65	20.4−20.9	2
8	6.52−10.75	0.50−1.00	19.4−20.5	7
9	24.00−30.08	0.75−1.00	19.8−20.5	3
10	11.88−14.84	0.59−0.67	19.5−20.4	5

diffraction analysis to ascertain that the crystalline type was consistently form V. Using the inferential model and the spectra collected from the solubility experiments, the solubility curve of form V was constructed.

The solubility curve for compound-X in toluene is shown in Figure 2. At 25 °C the solubility predicted by ATR FTIR spectroscopy and the chemometric method coincides with off-line measurement using high performance liquid chromatography (more HPLC comparisons are given below). The solubility points fit an exponential curve with high accuracy.

2. Compound-X in Single-Solvent (Toluene) System with Varying Temperature and Impurities. A similar procedure was applied using the crude compound-X. The presence of impurities results in colorization of the slurry. The typical purity level of crude compound-X is estimated to be 99.9% by the HPLC method. Consequently, the level of impurities is estimated to be below 0.1%. The presence of impurities did not affect the absorbance spectra of the compound-X−toluene solution. Therefore, the same inferential model built above was applied to measure the solubility. As shown in Figure 3, the presence of impurities slightly affected the solubility at high temperature**.** At lower temperatures, the solubility in pure and colored slurries appears to coincide. The solubility predictions from the chemometric model and ATR FTIR spectroscopy agree with HPLC data. The temperature profile followed for the measurement of the solubility was similar to that used for compound-X and the pure single-solvent system. For this case the crystals were determined to be of form V at the beginning and end of the experiment. Form V is the most thermodynamically stable crystal form.

3. Compound-X in Solvent (Toluene) and Antisolvent (*n*-Heptane) System with Varying Solvent Composition. Specified amounts of compound-X were dissolved in solvent (toluene). Once a spectrum was collected, a specified amount of antisolvent (*n*-heptane) was added, and a spectrum was collected again. This was repeated until crystals started to appear. Ten such solutions were prepared with various concentrations and vol % of toluene in the solvent−antisolvent system. The temperature ranged from 19 to 21 °C. Nitrogen at 20 °C was used as the background spectrum. Table 2 reports the number of spectra, the corresponding solution concentration range (g/kg of solution), the volume percent range of toluene in the solvent-antisolvent system, and the temperature range for each of the 10 solutions. The total number of spectra that were collected was 48.

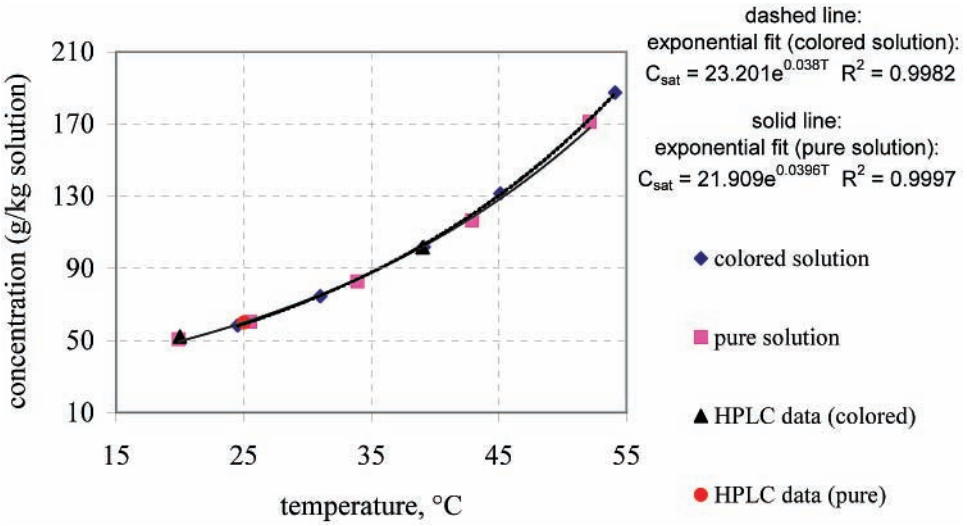

Figure 3. **Solubility curves of compound-X for pure and colored slurry and HPLC data.**

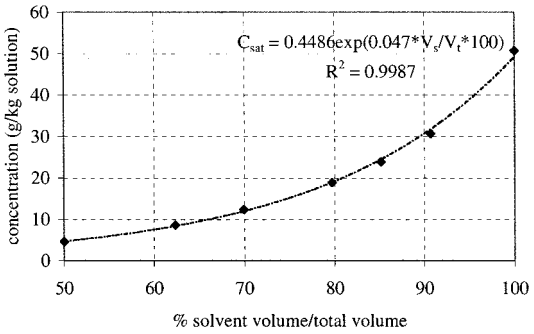

Figure 4. **Solubility curve of compound-X in solvent/antisolvent mixture.**

The above 48 spectra were used to construct the calibration curve, that is, the inferential model. The wavenumbers selected for the calibration were 799−1485 cm^{-1}. The chemometric model produced by the CPCR method gave an accuracy of compound-X concentration of ±7% or ±0.5 g/kg of solution at the 95% confidence level.

This inferential model predicts compound-X concentration from 50 to 100 vol % of toluene in toluene−*n*-heptane mixture at 20 °C. Some of the subdivisions gave poorer results. On the basis of examination of the data, the data in those calibration sets were not sufficiently dispersed throughout the range of variables.

After constructing the inferential model, solubility experiments were conducted. Appropriate amounts of toluene−heptane were mixed to create a binary solvent mixture of 50/50 volume solvent/volume antisolvent. Then an excess amount of compound-X was added to form a slurry. Spectra were collected once the spectra reach a plateau. After that, additional toluene was added to the slurry to form a new volume solvent/volume antisolvent mixture. The procedure was repeated until all crystals in the slurry dissolved. The measurements for 100 vol % toluene (no antisolvent present) were collected in the previous study. The solubility curve was constructed for a range of solvent composition of 50−100 vol % of toluene. This procedure ensures that the solubility curve is approached from undersaturation. Polymorph V is the most thermodynamically stable in the mixture of toluene and heptane. The crystalline type was determined to be form V at the beginning and at the end of the experiment.

Table 3. **Number of spectra, temperature, and concentration ranges for compound-Y**

	solution concentration range (g/kg solution)	solvent/total volume	temperature range (°C)	number of spectra
1	32.93−49.09	0.60−1.00	18.8−20.7	7
2	54.10−88.61	0.53−1.00	20.2−21.3	9
3	76.87−117.76	0.56−1.00	19.1−20.0	8
4	101.79−154.02	0.56−1.00	25.8−27.7	8
5	46.38−54.61	0.39−0.47	20.0−23.9	3
6	9.94−16.31	0.27−0.47	17.9−21.7	8
7	27.53−37.44	0.33−0.47	18.9−23.1	5
8	124.02−171.82	0.62−1.00	23.8−26.1	5
9	76.26−115.92	0.34−0.57	26.1−29.8	9
10	42.61−70.78	0.29−0.52	19.3−22.9	8
11	62.12−96.82	0.31−0.52	25.1−32.3	7
12	33.06−49.13	0.33−0.52	19.9−24.7	6
13	99.94−202.61	0.38−1.00	22.7−27.7	12
14	74.50−155.29	0.23−0.52	18.8−24.4	12

Using the chemometric model constructed for compound-X in toluene−*n*-heptane and the spectra collected from the solubility experiments, the solubility curve was constructed. The acquired points fit an exponential curve (see Figure 4).

4. Compound-Y in Solvent (10/1 Volume Ethyl Acetate/Volume 2-Propanol) and Antisolvent (*n*-Heptane) System with Varying Solvent Composition. A solvent mixture of 10/1 volume ethyl acetate/volume 2-propanol was prepared. Specified amounts of compound-Y were diluted in this solvent mixture. Once a spectrum was collected, a specified amount of *n*-heptane as antisolvent was added, and the spectrum was collected again. This was repeated until crystals started to appear. To enhance the collection of spectra for high concentrations of compound-Y and *n*-heptane, droplets of water were added to increase solubility. Fourteen such solutions were prepared with varying solute and solvent concentrations. As shown in Table 3, the total number of spectra that were collected was 107. Nitrogen at 20 °C was used as the background spectrum.

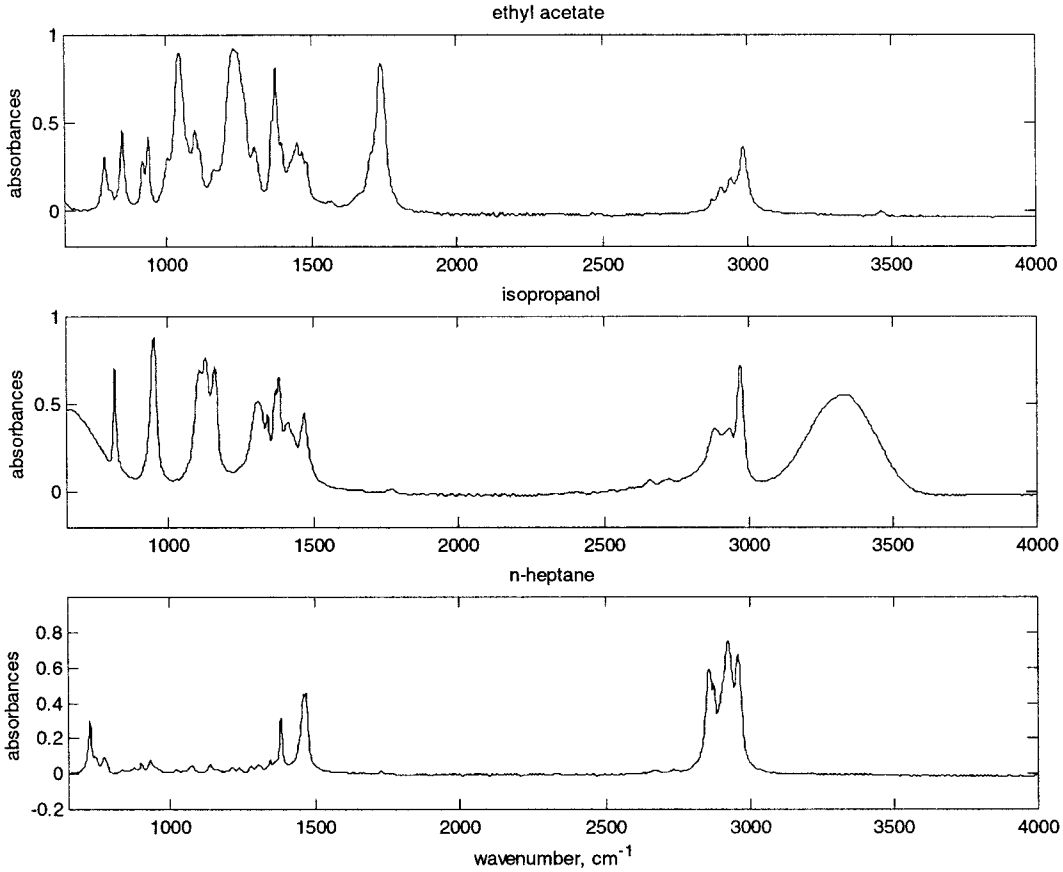

Figure 5. **Absorbance spectra of ethyl acetate, 2-propanol, and *n*-heptane at 20 °C.**

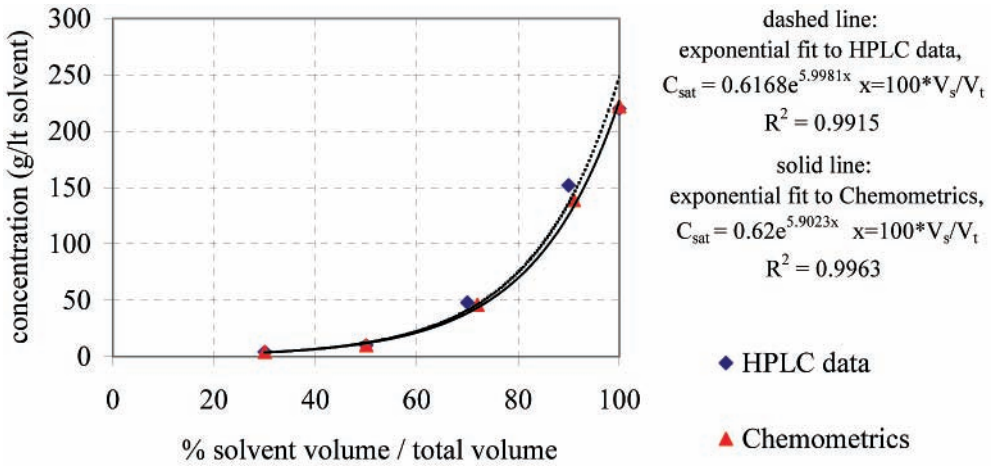

Figure 6. **Solubility curve of compound-Y at a water content of 2000 μg/mL.**

As shown in Figure 5 compounds of the system absorb strongly in a wide range of wavenumbers. To construct the inferential models, the absorbances for the wavenumber range of 982.7−1493 cm^{-1} gave more accurate predictions for ethyl acetate, 2-propanol, and *n*-heptane, whereas the range 982.7−1300 cm^{-1} gave more accurate results for the prediction of compound-Y. These wavenumber ranges were used in the subsequent results.

The CPCR method was used to construct the chemometric models for ethyl acetate, 2-propanol, and *n*-heptane. For this combination, the predicted volume ratios of ethyl acetate to 2-propanol were close to 10/1. The FPCR2 method was used to construct the chemometric model for compound-Y. Four inferential models were built at room temperature with a solvent composition ranging from 0 vol % to 70 vol % of *n*-heptane in *n*-heptane and 10/1 volume ethyl acetate/volume 2-propanol solvent mixture. For compound-Y, the accuracy of model is ±7% or +3 g/kg of solution at the 95% confidence. For ethyl acetate, the accuracy is ±1% or +3 g/kg of solution. For 2-propanol, the accuracy is ±5% or ±2 g/kg of solution. In addition, for *n*-heptane, the accuracy is ±7% or ±4 g/kg of solution.

For this particular system, the concentration of water affects the solubility of compound-Y significantly. The solubility curve was constructed at three different water contents (2000, 3000, and 3500 μg/mL) measured on solvent

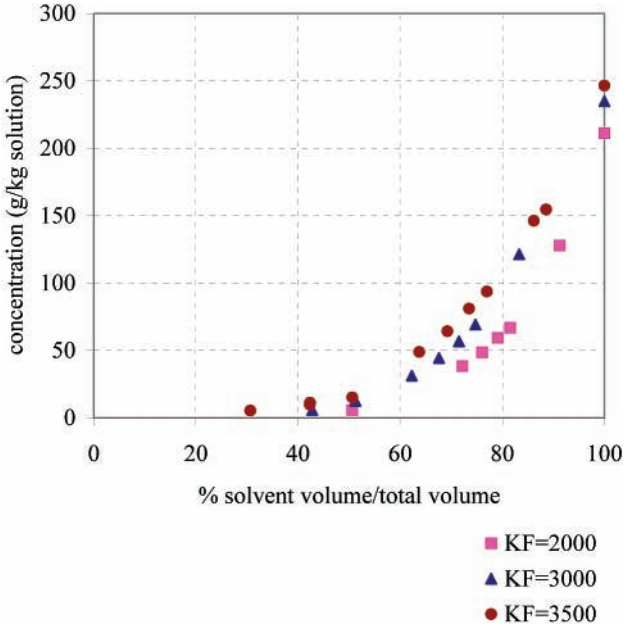

Figure 7. **Solubility predictions of compound-Y at water contents of 2000, 3000, and 3500 μg/mL.**

Table 4. **HPLC results and chemometrics prediction for solubility of compound-Y in g/L solvent with a water content of 2000 μg/mL**

HPLC		chemometrics	
vol % solvent in total volume	g/L solvent	vol % solvent in total volume	g/L solvent
30	4	30	4
50	10	50	10
70	48	72	46
90	152	91	139
100	220	100	222

basis with accuracy $\pm$150 μg/mL with 95% confidence. Experimentally, a solvent mixture of 10/1 volume ethyl acetate/volume 2-propanol was prepared, and the water content was adjusted to the desired value by adding droplets of water. Then an appropriate amount of *n*-heptane was added as antisolvent to create a solution of 30 vol % of solvent. Then an excess amount of compound-Y was added to produce a slurry. Spectra were collected once the spectra reach a plateau. Solvent was then added to the slurry to create a higher solvent composition, and a new spectrum was collected. The procedure continued until a high vol % of solvent was reached or until all the crystals of slurry were

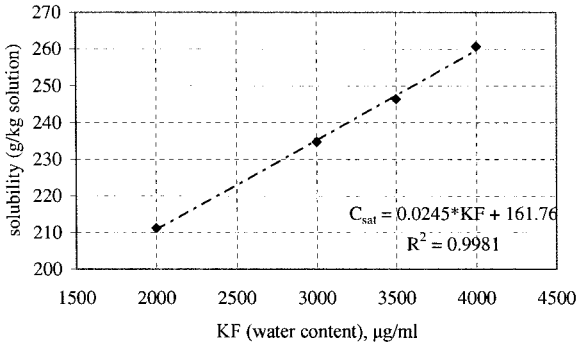

Figure 8. **Solubility of compound-Y at 100 vol % solvent (no antisolvent) as a function of water content (in μg of water/mL of solution).**

dissolved. The measurements for 100 vol % of solvent were taken separately. The solubility curve was constructed for a range of solvent compositions of 30−100 vol %. This procedure ensures that the solubility curve is approached from undersaturation.

In Table 4 the results for the HPLC and chemometrics methods are reported for a water content of 2000 μg/mL. As shown in Figure 6, chemometric predictions and HPLC are in good agreement, and both gave good fits on exponential curves. Figure 7 reports solubility curves for a water content of 2000, 3000, and 3500 μg/mL. Figure 8 reports the solubility of compound-Y in 100 vol % solvent (no antisolvent) as a function of water content for four different water concentrations. There is a linear relationship of solubility to water content for the investigated range of water concentrations.

Conclusions

For the first time, it has been demonstrated that inferential models constructed from ATR FTIR data and employing chemometrics techniques can predict solution concentration for pharmaceutical compounds in various multicomponent crystal slurries with high accuracy. The models cover a wide range of temperature and solvent composition for pure slurries or with the presence of impurities. Solubility data that are estimated using the inferential models are found to be in close agreement with the HPLC results.

Received for review December 4, 2001.

OP015516X