Journal of
Applied Crystallography

ISSN 0021-8898

Received 17 March 2006
Accepted 27 March 2006

Characterization of insulin microcrystals using powder diffraction and multivariate data analysis

Mathias Norrman,[a,b] Kenny Ståhl,[c] Gerd Schluckebier[a]* and Salam Al-Karadaghi[b]

[a]Diabetes Protein Engineering, Novo Nordisk A/S, Novo Alle 1, DK-2880 Bagsvaerd, Denmark, [b]Molecular Biophysics, Lund University, Box 124, SE-22100 Lund, Sweden, and [c]Department of Chemistry, Technical University of Denmark, DK-2800 Lyngby, Denmark. Correspondence e-mail: gesc@novonordisk.com

Twelve different microcrystalline insulin formulations were investigated by X-ray powder diffraction and were shown to have very characteristic patterns. Three of the formulations crystallize in the same crystal system, but have structural differences in the N-terminal B-chain of the insulin molecule. This difference was efficiently detected in the powder patterns. The sensitivity of the method makes it a valuable tool for characterization of microcrystalline samples. By use of principal-component analysis, the twelve different formulations originating from six different crystal systems were classified into nine separate clusters. The powder patterns of each cluster can now be used as 'fingerprints' for the different insulin polymorphs. The combination of X-ray powder diffraction and multivariate analysis, such as principal-component analysis, provides a rapid and effective tool for studying the influence of derivatives, additives, ions, pH *etc*., in the crystallization media.

1. Introduction

Injections of the blood-glucose regulating hormone insulin are used daily by patients worldwide suffering from diabetes. Insulin mediates the uptake of glucose from the blood system, and the inability to produce enough insulin or being insulin resistant requires supplemental insulin that primarily is administered by subcutaneous injections. In order to keep a constant glucose level and to respond to fluctuating levels after *e.g.* food intake, the action profile of the insulin pharmaceuticals needs to be both long-acting, keeping a basal level of insulin, and short-acting, to supply the circulating system with active insulin quickly. This is in most cases achieved by using two different pharmaceutical formulations. Since the 1920s, when insulin was first discovered as a drug (Banting & Best, 1922), a number of different formulations have been developed. The short-acting formulations often consist of free insulin in solution, while the basal, long- and intermediate-acting formulations are frequently composed of suspensions of microcrystals that are administered by subcutaneous injections. The insulin molecule comprises two polypeptide chains, A and B, with 21 and 30 amino acids, respectively. The chains are linked by two interchain disulfide bridges, and there is one intrachain disulfide bond in the A-chain. Insulin in zinc-free solutions exists as a dimer, which associates into hexamers in the presence of divalent metal ions, like zinc (Schlichtkrull, 1958). When the suspension is injected into the subcutis, dissolution of the crystal is rate-limiting for absorption into the bloodstream. Hence the crystalline nature and the size of the crystals both contribute to the duration time of the suspension formulation (Brange, 1987). Once the crystals have disbanded, the constituent hexamers, dimers and monomers diffuse through the capillaries. Due to their different size, hexamers have a longer diffusion time than dimers and monomers. Other factors that may have an impact on the profiles of action are the crystal form, morphology (Pechenov *et al.*, 2004), crystal packing and composition of crystals, which can be affected by the presence of additives and ligands, such as zinc ions and phenolic molecules.

All the factors which control stability, bioavailability and delivery of insulin are extensively studied in order to design new insulin pharmaceuticals with desired properties. Careful chemical and physical characterization of insulin microcrystals is also important for the control of the homogeneity of batches at the production line. The hexameric insulin in the zinc-containing formulations can exist in three different conformations, called T_6, $T_3R_3^f$ and R_6 (Kaarsholm *et al.*, 1989), which refer to the folding of the N-terminal part of the B-chains (Fig. 1). In the T_6 form, residues B1 to B8 are found in an extended configuration (Baker *et al.*, 1988; Smith *et al.*, 2003; Smith & Blessing, 2003), whereas in the R_6 form the same residues are in a helical conformation (Derewenda *et al.*, 1989; Smith & Dodson, 1992; Smith *et al.*, 2000). This, together with the already existing helical conformation of residues B9 to B19, creates a long continuous α-helix between residues B1 and B19. In the $T_3R_3^f$ form, three insulin molecules are found in T conformation, while the other three molecules are in a 'frayed' R conformation, where the first three residues in the

B-chain are found in a non-helical conformation (Ciszak & Smith, 1994; Ciszak *et al.*, 1995; Smith & Ciszak, 1994; Whittingham *et al.*, 1995).

The first stable protracted preparation of insulin, the NPH (Neutral Protamine Hagedorn) was introduced in 1946 (Krayenbuhl & Rosenberg, 1946). The insulin–zinc solution was cocrystallized with the basic peptide protamine, which consists mainly of arginine residues. This polypeptide reduces insulin solubility. Each hexamer contained two zinc atoms and one protamine peptide and was crystallized at pH 7.3 in the tetragonal crystal system with space group $P4_32_12$ (Balschmidt *et al.*, 1991). The commercial products Penmix30 (human insulin) and Novomix30 (ProB28Asp) consist of a mixture of soluble and crystallized NPH insulin in the ratio of 30/70. The Protaphan formulation consists of 100% crystals from pig insulin (ThrB30Ala). Another type of insulin formulation with an even more prolonged action profile is often referred to as Lente insulins (Hallas-Møller, 1956; Hallas-Møller *et al.*, 1951). They consist of rhombohedral crystals of space group $R3$ and contain hexameric insulin with two zinc atoms at the threefold axis and one phenolic derivative per monomer. The commercial products Ultratard and Ultralente consist of 100% crystalline human insulin. A third type, the Lente product, consists of one third amorphous pig insulin and two thirds crystalline bovine insulin.

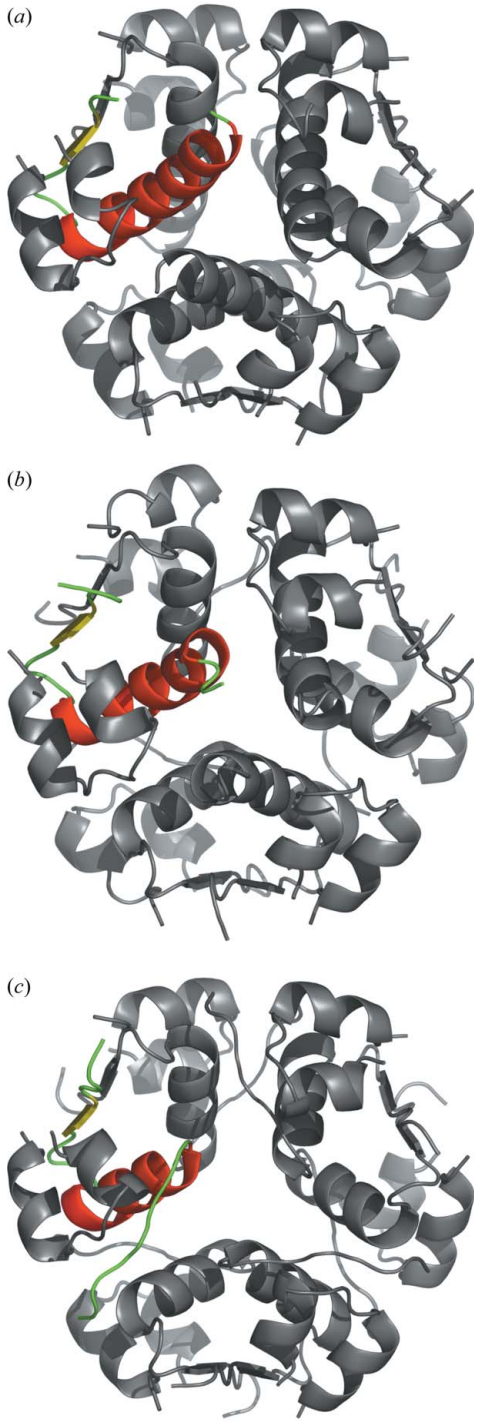


Figure 1
Three different conformations of B-chain folding in insulin hexamers. One B-chain in each hexamer is highlighted and coloured according to secondary structure. (*a*) R_6 form, residues 1–8 in the B-chain are in helical conformation, which together with residues B9–B19 creates a continuous helical fragment. (*b*) $T_3R_3^f$. Residues B1–B3 have an extended conformation, while residues B4–B19 are helical. (*c*) T_6. Residues B1–B8 are in an extended conformation.

In the manufacturing process, cubic insulin crystals are routinely used in the purification and capture steps. The space group for these crystals is $I2_13$ with one monomer in the asymmetric unit. These cubic crystals are grown by including 1 *M* NaCl at pH 6.5, in zinc-free crystallization media without phenolic derivatives (Harding *et al.*, 1966).

In this study, we use X-ray powder diffraction as the method for the analysis of several different microcrystalline insulin formulations, both from commercial products and from in-house-developed preparations. The microcrystals range from 10–25 µm in size, making single-crystal structure determination very tedious, if not impossible. X-ray powder diffraction is mostly used to study small-molecule structures, but has been shown to be applicable also to smaller proteins. Powder data from microcrystals of a variant of human insulin (LysB28 ProB29) have been studied and compared with simulated and recorded patterns of rhombohedral insulins (Richards *et al.*, 1999). In addition, metmyoglobin and rhombohedral $T_3R_3^f$ insulin have been analysed using high-resolution powder diffraction and molecular replacement techniques (Von Dreele, 1999; Von Dreele *et al.*, 2000). By using powder diffraction, the same authors have also studied binding of *N*-acetylglucosamine to hen egg-white lysozyme (Von Dreele, 2001, 2005). Turkey egg-white lysozyme has been used as a model system to demonstrate the use of high-resolution powder diffraction and molecular replacement techniques in the study of small structural variations in protein molecules (Margiolaki *et al.*, 2005; Basso *et al.*, 2005). In the present study, we have analysed 12 insulin products or formulations. The results show that medium-resolution X-ray powder diffraction in combination with multivariate data analysis for fingerprinting can be used for the comparison of similarities and differences between preparations of insulin microcrystals. This demonstrates that the method can be used to distinguish between different crystal systems and to assess homogeneity of different batches or preparations of insulin. It may also be used for finding novel insulin formulations and in the study of

Table 1
Crystallographic properties of the samples used in the study.

Crystal	λ (Å)† / beamline	Trade name	Crystal system	Space group	Seq. origin‡	B-chain configuration	Unit cell				PDB ref.§
							a (Å)	*b* (Å)	*c* (Å)	*β* (°)	
A	0.97 / 911-3		Monoclinic	$P2_1$	H	R_6	61.3	61.7	47.5	111.3	1ev6[(1)]
B	1.00 / 911-2		Orthorhombic	$C222_1$	H	R_6	58.9	219.4	223.7		In-house database
C	1.00 / 911-2		Cubic	$I2_13$	H	T_6	78.9	78.9	78.9		1aph[(2)]
D	0.97 / 911-3	Detemir	Rhombohedral	*R*3	H	R_6	78.9	78.9	39.5		1ev3[(3)]
E	0.97 / 911-3		Rhombohedral	*R*3	H	$T_3R_3^f$	80.6	80.6	37.8		1trz[(4)]
F	0.969 / 711	Ultralente	Rhombohedral	*R*3	H	T_6	81.3	81.3	33.7		1mso[(5)]
G	0.969/ 711	Ultratard			H		82.5	82.5	34.0		4ins[(6)]
H	0.969 / 711	Lente			B, P						
I	0.969 / 711	Penmix30	Tetragonal	$P4_32_12$	H	R_6	62.9	62.9	85.9		In-house database
J	1.00 / 911-2	Novomix30	Tetragonal	$P4_32_12$	H B28Asp	R_6	62.8	62.8	86.9		In-house database
K	0.969 / 711	Protaphan	Tetragonal	$P4_32_12$	P	R_6	62.9	62.9	85.9		7ins[(7)]
X	0.97 / 911-2		Unknown	Unknown	H	Unknown					

† Wavelength used during data collection. ‡ H = human, B = bovine, P = porcine. § Coordinate files used for simulated powder patterns. References: (1) Smith *et al.* (2000); (2) Gursky *et al.* (1992); (3) Smith *et al.* (2000); (4) Ciszak & Smith (1994); (5) Smith *et al.* (2003); (6) Baker *et al.* (1988); (7) Balschmidt *et al.* (1991).

Table 2
Crystallization conditions for the samples used in the study.

	A	B	C	D	E	F	G	H	I	J	K
Space group	$P2_1$	$C222_1$	$I2_13$	*R*3	*R*3	*R*3	*R*3	*R*3	$P4_32_12$	$P4_32_12$	$P4_32_12$
Insulin (mg ml^{-1})	5.2	3.5	10	14	3.8	3.8	3.8	1.5	3.8	3.8	1.5
Zn/hexamer	2.3	2.3		2.5	4	22	22	22	3	3	3
Phenol derivative (m*M*)†	20[(1)]	25[(1)]		19[(3)]/19[(4)]		65[(2)]	65[(2)]	65[(2)]	7[(3)]/14[(4)]	7[(3)]/14[(4)]	7[(3)]/14[(4)]
NaCl (*M*)		1.0	1.0	0.02	0.3	0.12	0.12	0.12			
Na acetate (*M*)					0.01	0.01	0.01	0.01			
Na citrate (*M*)				0.11							
Na_2HPO_4 (*M*)	0.48	0.05	0.04						0.013	0.013	0.013
Urea (*M*)		1.1									
Tris (*M*)				0.14							
Protamine									Added in isophane ratio‡		
pH	7.3	6.7	7.2	8.15	5.5	7.4	7.4	7.4	7.3	7.3	7.3

† Phenol derivatives: (1) resorcinol (benzene-1,3-diol); (2) methyl *p*-hydroxybenzoate (methyl 4-hydroxybenzoate); (3) phenol; (4) *m*-cresol (3-methylphenol). ‡ Protamine is added in isophane ratio (Balschmidt *et al.*, 1991; Krayenbuhl & Rosenberg, 1946).

the role of various additives, ions, *etc*., in crystallization. The insulin formulations used in this study are given in Table 1, along with the naming convention (A, B, C . . .) that will be used throughout this paper when referring to the different crystal types.

2. Materials and methods

2.1. Insulin samples

Ultratard, Ultralente, Lente, Detemir, Penmix30, Novomix30 and Protaphan formulations were obtained from Novo Nordisk A/S. Other microcrystals were prepared by batch crystallization. In-house-developed NPH-like preparations were crystallized according to Balschmidt *et al.* (1991). The rhombohedral crystals with $T_3R_3^f$ configuration were crystallized in batch mode, following the same procedure as in the first step of Ultralente crystallization (Hallas-Møller, 1956; Hallas-Møller *et al.*, 1951). A novel type of human insulin crystals was prepared in 1.1 *M* Urea, 1 *M* NaCl, 25 m*M* resorcinol at pH 6.7, with 2.3 Zn per hexamer, which yielded orthorhombic $C222_1$ crystals (a = 59, b = 219, c = 223 Å) with three hexamers in the asymmetric unit and with R_6 configuration of the B-chain. Details on crystallization and the three-dimensional structure will be reported elsewhere (Norrman *et al.*, in preparation) The crystals were about 0.15 mm in size and were therefore crushed before powder X-ray diffraction analysis. An insulin polymorph (X), with unknown crystallographic properties, was obtained by a propriety in-house formulation screen. Crystallization conditions for all other formulations used in this study are summarized in Table 2.

2.2. Sample preparation and powder diffraction

The microcrystal suspensions were transferred to a bottom-capped glass capillary (Hampton Research, USA) with an outer diameter of 0.7 mm and centrifuged at 1500 g for 15 min to pack the crystals in the bottom of the capillary. The capillaries were sealed and mounted on a goniometer head. Powder data at room temperature were collected both in-house, using a rotating anode generator (Rigaku RU200, Osmic mirrors, Cu $K\alpha$ radiation, λ = 1.5418 Å) with a Mar345 imaging plate, and at the Max-lab synchrotron (Lund, Sweden), beamlines 711 (Cerenius *et al.*, 2000), 911-2 and 911-3 (Mammen *et al.*, 2002), on different occasions (different wavelengths), using a CCD

detector. Typical exposure times were 1 h ($\Delta\varphi = 360^\circ$) for in-house data collections and 1 min ($\Delta\varphi = 60^\circ$) for synchrotron data.

2.3. Data analysis

The experimental powder patterns were first analysed using the *Datasqueeze* software (http://www.datasqueezesoftware.com), by which the intensities in the 2θ range 0.9–10° were integrated by summation of the intensities in the χ region 0–360°. The resulting plots of the powder profiles were saved in xy-files (intensity *versus* 2θ) in ASCII format and imported into the *WinPrep* program (Ståhl, in-house program) for background correction and smoothing. Since the synchrotron data were collected at different beamlines and with different wavelengths (Table 1), the 2θ values were converted into *d*-values in order to align different data sets. After alignment, the *d*-values were re-converted to 2θ using a primary data set with $\lambda = 0.969$ Å as reference. This wavelength serves as reference for all analyses and plots made in this study. All intensities were normalized against the total intensity using in-house software. A file containing normalized intensity data as a function of 2θ with an increment of 0.009° in 2θ was saved in an in-house-developed database. No peak fitting was applied to determine the peak centres; instead the peak maximum was used. The processed powder diffraction data have been deposited with the International Center for Diffraction Data (ICDD; http://www.icdd.com).

2.3.1. Principal-component analysis. For easy and objective visualization of the samples, the powder patterns were analysed by principal-component analysis (PCA) (Wold *et al.*, 1987) using *Simca-P*+ software (Umetrics AB, Umeå, Sweden; http://www.umetrics.com). PCA is a projection method to visualize complex data by reducing the dimensionality in a data set, typically into two or three dimensions. The data consist of a matrix with *N* rows (observations) and *K* columns (variables). The number of dimensions in the data set at the starting point is equal to the number of columns (*K*). Dimensionality is reduced by finding a plane in the multi-dimensional space with the largest variation. This plane is referred to as a principal component (PC). Once the first PC is found, another one, orthogonal to the first, is searched. When a number of components are found, all of which being orthogonal to each other, the observations are projected into a new coordinate system, where the principal components form the axes. Plots based on this coordinate system are referred to as score plots. The score plots can be used to reveal clustering (grouping) of the samples and to detect outliers. For an introduction to PCA, see the work of Wold *et al.* (1987).

In this study, all intensity data points in the 2θ range 0.9–6.0° (step size 0.009°) were used and scaled by unit variance (UV), prior to the PCA, thereby weighting all peaks equally. Sample similarities were analysed by loading a table with the samples (observations) in rows and their intensity data points as a function of 2θ (variables) in columns. The sample distribution was analysed in score plots with the observations projected in two or three dimensions.

2.4. Calculation of powder patterns

Simulated powder patterns were calculated from atomic coordinates for the insulin polymorphs where single-crystal structures were available. The patterns were calculated using the *WinPrep* program and compared with experimental data for confirmation of crystal system. A Lorentz factor of $1/\sin\theta$ (Warren, 1990) and a polarization factor of $(1 + \cos^2 2\theta)$ were applied to the calculated intensities. Both corrections were applied to data at the originally measured wavelength, and the patterns were subsequently recalculated to a common wavelength of 0.969 Å. Profile parameters for the full width at half-maximum and pseudo-Voigt (γ) factor (the pseudo-Voigt function is a linear combination of a Gaussian and a Lorentzian function, where γ describes the weighting between the two) were set to 0.07 and 0.5, respectively. For comparison of experimental and calculated patterns, the patterns were normalized against the total intensity in the 2θ region from 2.5 to 10°. The experimental powder data were collected at room temperature, while the single-crystal data corresponding to crystals A, D and F were collected at 100 K. Examples of cryo-cooled induced changes in unit-cell dimensions are well documented for insulin and other systems (Smith *et al.*, 2003; Halle, 2004). To illustrate the effect of temperature-induced changes of the cell constants on the powder pattern, an insulin structure obtained at room temperature, PDB code 4ins (pig insulin) (Baker *et al.*, 1988), with slightly larger unit-cell dimensions ($a = 82.5$, $b = 82.5$, $c = 34.0$ Å) was used as an additional reference for the F crystals.

3. Results

All samples of crystalline insulin gave rise to powder diffraction patterns using standard protein crystallographic equipment. Since patterns obtained with synchrotron radiation generally had sharper and better resolved peaks, they will be used in the following discussion. Representative samples of raw data and the resulting intensity *versus* 2θ plots are shown in Fig. 2. Clearly, diffraction patterns for different insulin polymorphs had distinct peaks in the low-2θ region (0.9° to ~6°), making powder diffraction a method of choice when comparing a large numbers of microcrystalline samples.

Visual comparison of the plots in Fig. 2 shows that crystals which belong to the same crystal system and which have the same type of structure have very similar powder patterns. However, even small differences in protein structure result in detectable differences in the powder patterns. The NPH crystals I, J and K crystallize in the same crystal system (tetragonal $P4_32_12$), but differ in that J has a ProB28Asp mutation, which introduces an additional negative charge, K crystals consist of 100% pig insulin (ThrB30Ala), and the I crystals are from human insulin. As shown in Fig. 2, the overall patterns from this group of crystals have a high degree of similarity. The I and K crystals are the most similar with a good match in the low-2θ region. The major difference is an additional peak at $2\theta = 4.1^\circ$ in the K pattern (marked with an arrow in Fig. 2*c*) that is not found in the I pattern. In the pattern from

the J crystals, peak positions are shifted relative to the peak positions of the I and K crystals in the whole region. The introduction of an additional negative charge leads to a higher proportion of the cocrystallized basic protamine peptide being bound to the insulin (Balschmidt, 1996) resulting in slightly larger unit-cell constants (Table 1) and consequently a shift in peak positions. The different unit-cell content also leads to different diffraction intensities (peak heights).

Since for the F, G and H crystals (rhombohedral with T_6 conformation), peak positions are essentially identical, and only small differences in some peak intensities could be seen, the F crystals will be used in the following discussion when referring to this group of crystals. The overall patterns from the F (T_6), D and E crystals (rhombohedral with R_6 and $T_3R_3^f$ conformation, respectively), are shown in Fig. 3. As seen from the figure, similar peaks in the three patterns are generally shifted by less than 0.15° in 2θ. The region with the largest differences is found between 2θ values of 3.95 and 4.35°, where all groups have a high-intensity peak, but its position is clearly different: for the D crystals the peak maximum is at 4.01°, for E it is at 4.13° and for F at 4.31°. Among the differences, there is also a large peak at 2θ = 1.36° in the F pattern, which is smaller in the D and E patterns, and also shifted by +0.05° in the D sample. An additional peak, at 1.62° and 1.72° respectively in the D and E patterns, is missing in the F crystal pattern. The shifts in peak positions are most likely due to

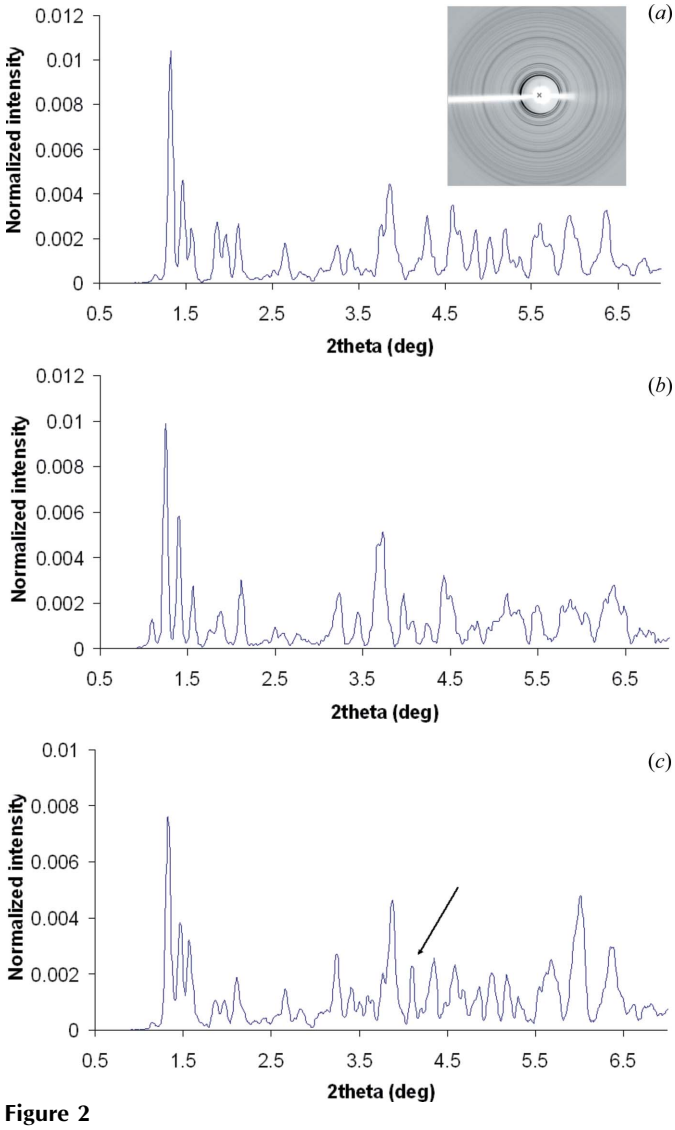


Figure 2
Comparison of diffraction patterns from the I (*a*), J (*b*) and K (*c*) crystals. The I and K crystals have a high degree of similarity (human and pig insulin, respectively). A slight shift in 2θ is seen for the J crystals, probably the result of slightly larger unit-cell constants (Table 1), caused by small structural differences due to a B28Asp mutation in the J crystals that gives an increased binding of the cocrystallized protamine peptide. For each sample, the intensity is normalized against total intensity. Inset is the diffraction pattern of the I crystals, obtained on a CCD detector at the beamline 711 (Max-lab, Sweden). The arrow in (*c*) indicates the extra peak found in the K crystals, but not in the I crystals.

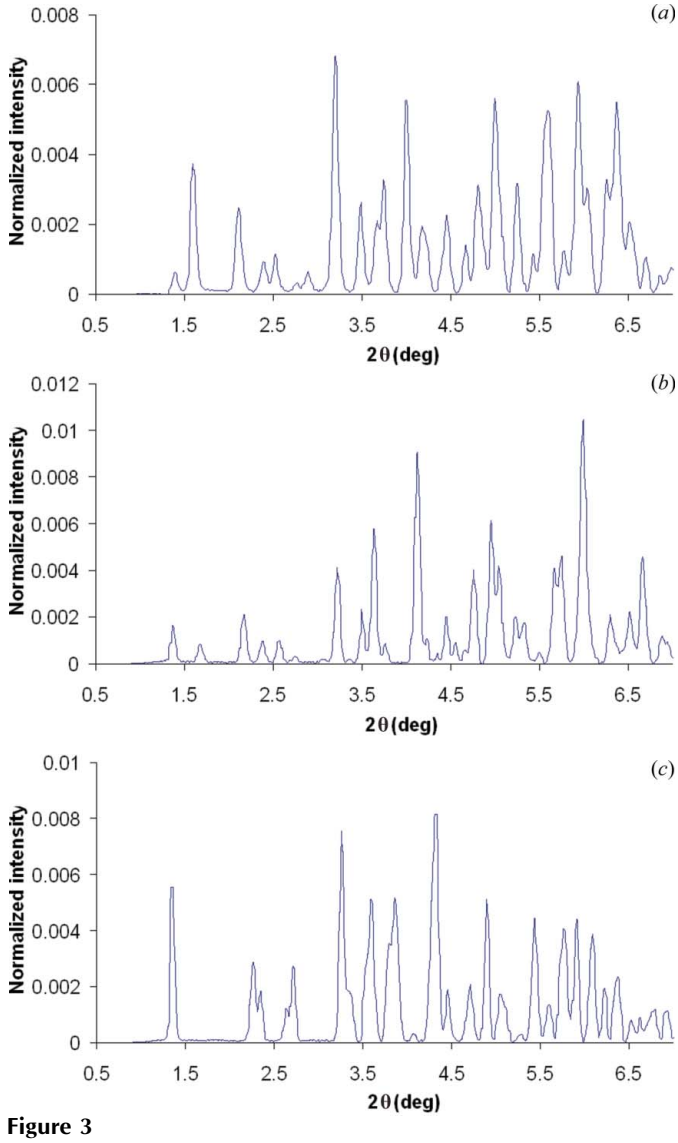
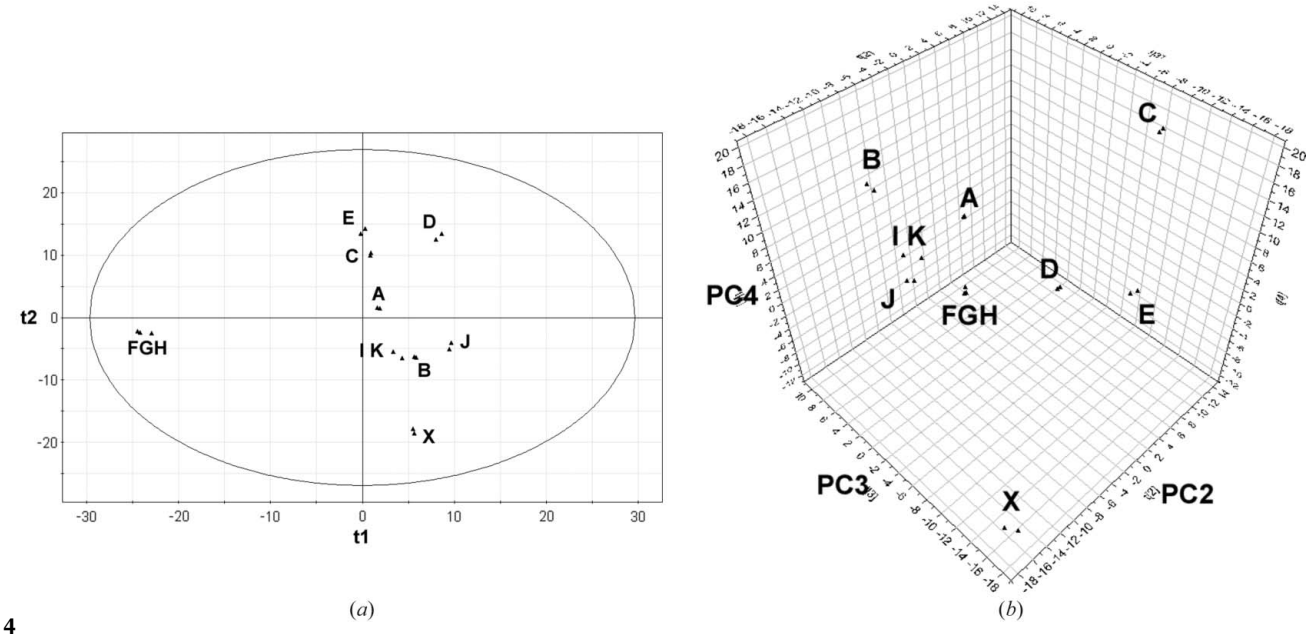


Figure 3
Normalized intensity *versus* 2θ for three types of rhombohedral crystals, D (*a*), E (*b*) and F (*c*). The three types have structural differences in the arrangement of the N-terminal residues of the B-chain (R_6, $T_3R_3^f$ and T_6 conformation, respectively). For each sample, the intensity is normalized against total intensity. The region with the largest difference is found between 3.95° and 4.35°, but there is a small shift in peak position throughout the whole 2θ region.



(*a*) (*b*)

Figure 4
(*a*) Score plot from the principal-component analysis with the two largest components (PC1, horizontal, and PC2, vertical). The scores t_1 and t_2 are new variables computed as linear combinations of all variables (intensities as a function of 2θ) and aim at describing as much of the original variation as possible without losing information. The first component (PC1) is mainly related to the variation of the F, G and H on the left side and the D and J crystals, on the right. The vertical plane (PC2), is related to the differences of the E, D and X crystals. The ellipse indicates a 95% confidence interval. Samples outside this region are possible outliers. The first two principal components (PC1 and PC2) are not enough to separate the I, K crystals fully from the B crystals. The three-dimensional plot in (*b*) is plotted with PC2, PC3 and PC4 and shows that those crystals are better separated by the fourth component. In total, nine separate groups can be seen where each group or cluster contains crystals from the same crystal system and/or with the same structural arrangement. Two samples of each crystal type have been included, except for the F crystals where three samples (F, G and H), originating from different formulations, are included to show their high degree of similarity. All sample data were collected at synchrotron beamlines (Table 1).

structural differences in the N-terminal part of the B-chain causing differences in the cell constants (Fig. 1 and Table 1). A comparison of the powder patterns of the D and A samples (A crystals being monoclinic with R_6 conformation) (data not shown), *i.e.* two different crystals with the same B-chain conformation, shows large differences in the peak positions, a clear indication that the A sample belongs to a different crystal system.

3.1. Principal-component analysis of powder patterns

Visual analysis of the powder patterns as described above is possible for a small number of samples, but as the number increases, the complexity becomes high and the procedure is very time-consuming. It was in the interest of this study to identify a method that could facilitate analyses and interpretation of the powder patterns from a larger number of microcrystal suspensions. We therefore utilized principal-component analysis to obtain a visual representation of the relationships and similarities of the samples. A similar method is incorporated into the commercial software *PolySNAP* (Bruker) (Barr *et al.*, 2004*a*,*b*) for the analysis of small-molecule diffraction data.

The basic objective in PCA is to reduce the dimensionality (number of variables) of the data set from several hundreds to two or three principal components, retaining most of the original variability in the data, *i.e.* without losing information. A typical PCA score plot is shown in Fig. 4(*a*). It is plotted in two dimensions using the two principal components (PC1 and PC2) which account for the largest variations in the data set: 22% and 18%, respectively. A two-dimensional projection of the results of the analysis considerably facilitates the comparison of intensity patterns from different crystal samples. The positions of the data points, corresponding to each sample, provide an overview of the relationship between samples or groups of samples. As seen in Fig. 4(*a*), some of the samples are clearly grouped into clusters. The clustering indicates a high similarity within each group, and a true difference between groups. The largest separation is found along the horizontal plane (PC1), with the F crystals in the left part of the plot and the D and J crystals in the right. This indicates that the first component primarily reflects the differences between these crystal types. On the other hand, the differences between the E, D and X crystals are marked along the vertical plane (PC2). The relative shifts in peak position as observed in the powder patterns of the three rhombohedral D, E and F crystals with characteristic differences in B-chain conformation (R_6, $T_3R_3^f$ and T_6 conformation), have a large impact on the distribution of their PCA scores in the plot. The D and F crystals are well separated along the PC1 axis, while the PCA score for the third rhombohedral crystal type, E, is found closer to the D form. It can also be seen in Fig. 4(*a*) that some crystal types are gathered in the central part of the plot (B and I, K), indicating that the first two principal components (PC1 and PC2) are unable to separate all samples. Since the first component is dominated by the F crystals, the three-

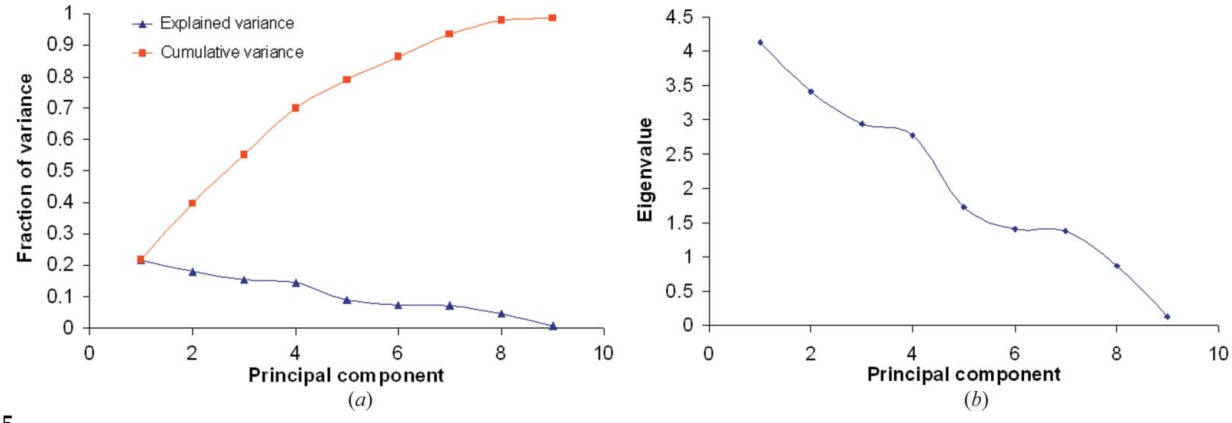


Figure 5
(*a*) Amount of explained variance in the data for each principal component. The lower line shows the fraction of explained variance, while the upper line shows the cumulative explained variance. The first four components explain in total 70% of the variability in the data set. (*b*) Plot of the eigenvalues for each component. A component is generally considered significant if its eigenvalue is higher than 2. Using this criterion, the first four components are significant and thus describe meaningful data.

dimensional plot in Fig. 4(*b*) is plotted with PC2, PC3 and PC4, accounting for 18%, 15.5% and 14.5% of the variation, respectively. The B sample is now distinguished from the I, K samples along the fourth component. In this representation, different crystal systems and/or structural arrangements are well separated, facilitating the identification of novel polymorphs. Fig. 5(*a*) shows the fraction of the total variance explained by each principal component along with the accumulated explained variation. The amount of data variability explained by the first four components is in total 70%. The number of components to include in an analysis is typically decided by the eigenvalue for each component. A component with an eigenvalue above 2 is considered significant, which in this case indicates that the first four components describe real and meaningful variability in the data (Fig. 5*b*). The predictability (fraction of the total variation that can be predicted) with these four components is moderate (37%). The PCA is therefore only used for visualization and for providing an overview of the sample distribution.

The relationships between observations (crystal samples) and the variables (intensities as a function of 2θ) can be visualized using a so-called loading plot. Such a plot shows the most important variables for a sample in the score plot. Fig. 6 shows a one-dimensional loading plot for the first principal component (PC1), coloured in orange. The 2θ values of the major positive peaks in the loading plot are related to the characteristic peak positions of the samples on the positive side of the PC1 axis in Fig. 4(*a*). Likewise, the negative peaks coincide with the samples on the negative side of the PC1 axis. PC1 is heavily dominated by the F crystals (located far to the left). The loading plot is combined with the powder pattern of the F crystals (blue) to illustrate this relationship. The peaks from the F crystals superimpose well on the loading line plot. The positive loading plot peaks originate from the samples located on the right side of the plot in Fig. 4(*a*), primarily D and J. Thus, from the PCA score plot and its loading plot, it is possible to deduce the primary variables (2θ values) determining the positions of the samples in the score plots. Another example is provided by the I and J samples. The analysis of the

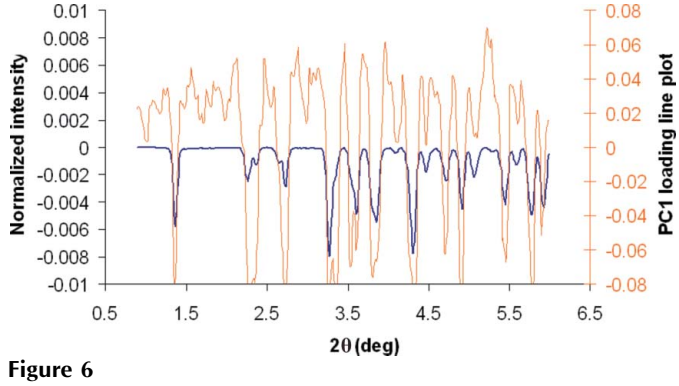


Figure 6
Loading plot for the first principal component, PC1, of the principal-component analysis (orange). A loading plot describes the relationships between the observations (crystal samples) and the variables (intensities). A plot of the loading vectors indicates which of the variables are important, and corresponds to the directions of the samples in the score plot, *i.e.* variables with positive peaks in the loading plot are important for samples on the positive side of its component in a score plot. The major positive peaks in the loading plot are thus important for the samples on the right side of the score plot in Fig. 4(*a*). Likewise, the negative peaks are important for the samples on the left part of the score plot. The powder pattern of the F crystals is coloured in blue. The normalized intensity of the F crystals is here put on a negative scale for clarity. The negative peaks of the loading plot coincide very strongly with the peaks of the F crystals, indicating that the PC1 is dominated by the F crystals. The positive loading peaks coincide with the peaks of the crystals located on the right side of the score plot in Fig. 4(*a*) (D, J and B).

position of the I and J crystals in the PCA score plot, as seen in Fig. 4(*a*), shows that the separation is dominated by the first component, PC1. A contribution plot (not shown) was used to deduce the dominating 2θ values, responsible for the observed separation; (I crystals/J crystals) 1.34°/1.22°, 3.88°/3.68° and 4.62°/4.42°. These peaks coincide with the 2θ positions of major peaks in the powder patterns.

3.2. Comparison of experimental and calculated powder diffraction patterns

Powder patterns of the known insulin polymorphs were calculated from coordinate files using the *WinPrep* program.

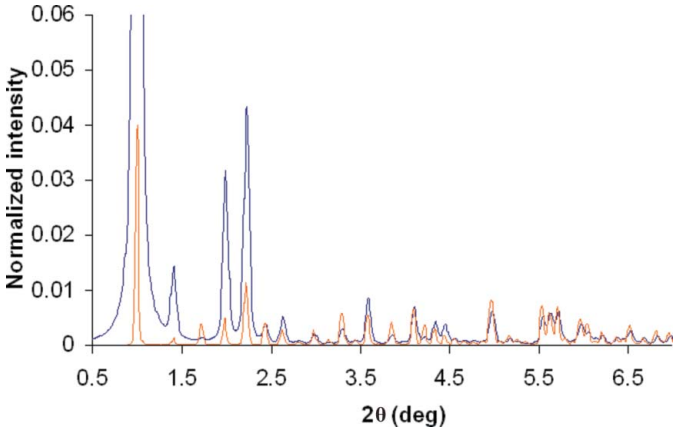


Figure 7
Comparison of a simulated (dark blue) cubic powder pattern with experimental data (orange) obtained from the cubic C crystals. The peak positions have good agreement between simulated and experimental data, but the intensity difference is large in the low-2θ region. Normalization against the total intensity in the whole 2θ range would suppress the lower intensity peaks. In order to limit the influence of the low-2θ peaks, the normalization was performed against the total intensity in the 2θ range 2.5–10°.

Visual comparison of the profiles shows that the peak positions of the simulated cubic type crystals superimpose almost exactly on the experimental patterns (Fig. 7). In some of the other crystal types, either a few peak positions or some peak intensities are skewed. The overall similarity of the peak positions is, however, good enough to conclude on the agreement between the predicted and experimental patterns. Differences between the prediction for the rhombohedral crystals D, E and F reflect the same differences which were seen in the experimental patterns between the R_6, $T_3R_3^f$ and T_6 conformations (Figs. 8*a*-8*d*). Although some of the peak intensities differ, the majority of the peak positions are the same in the simulated and experimental data. The largest peak position deviations are found between the simulated and experimental D and F crystals (Figs. 8*a* and 8*c*). In the 2θ region higher than 4.0°, the simulated pattern is slightly shifted to the right. The experimental powder data were collected at room temperature, while the single-crystal data were collected at 100 K. As an additional reference for the F crystals, the pig insulin 4ins (Baker *et al.*, 1988), obtained at room temperature and with slightly larger unit-cell dimensions (a = 82.5, b = 82.5,

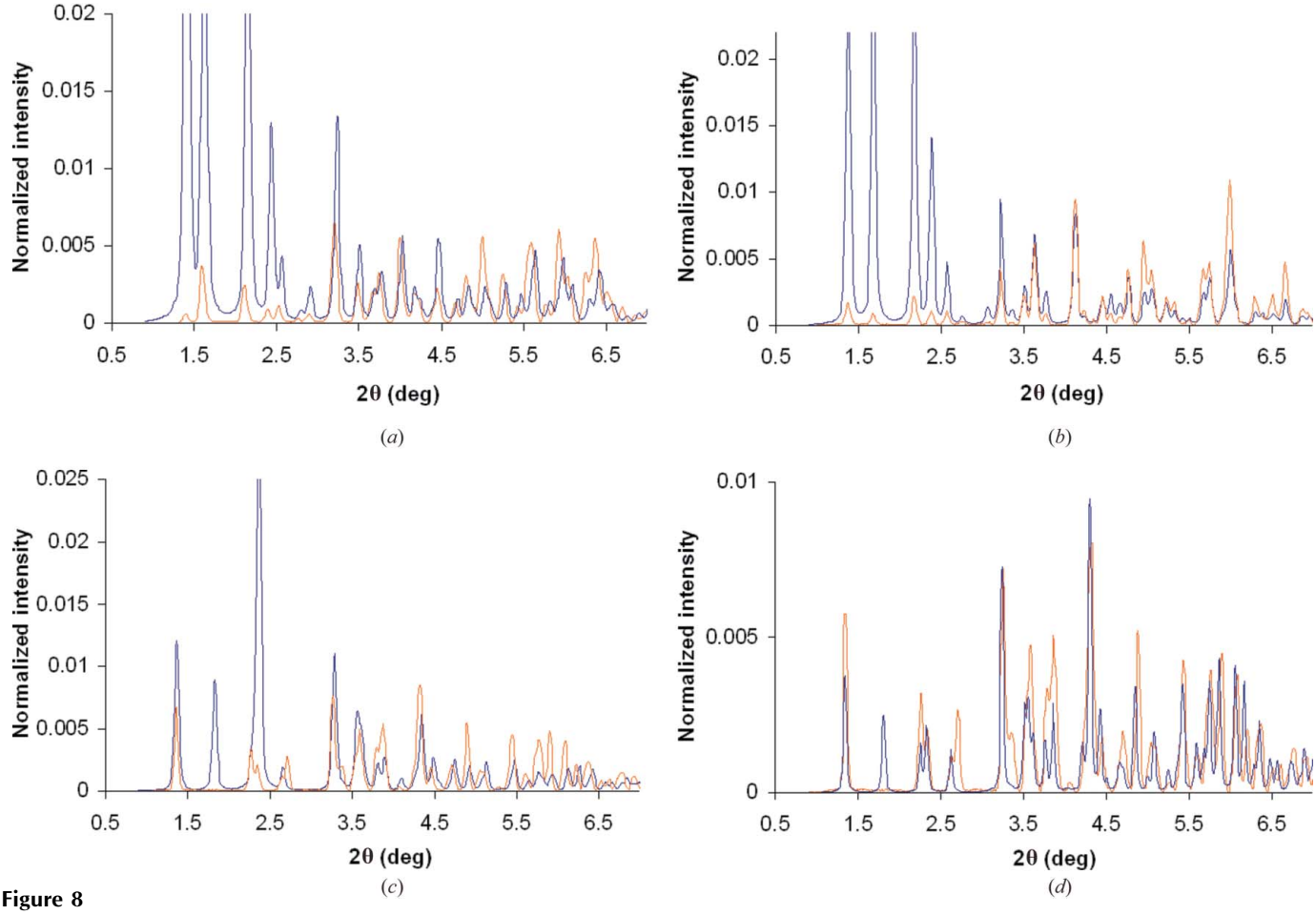


Figure 8
(*a*)–(*d*) Comparison between simulated and experimental powder patterns of the three rhombohedral crystal types D, E and F. The three types differ in the length of the c axis and their N-terminal B-chain conformation (R_6, $T_3R_3^f$ and T_6). Simulated patterns are coloured in dark blue, experimental in orange. (*a*) D crystals, (*b*) E crystals, (*c*) F crystals compared with 1mso (human insulin collected at 100 K), (*d*) F crystals compared with 4ins (pig insulin collected at room temperature). The room-temperature structure has a better match of the peak positions, indicating that the cell dimensions of the F crystals are more similar to this one. Normalization was performed as explained in Fig. 7. Although the intensities differ, the peak positions have a good overall match between simulated and experimental data.

c = 34.0 Å) was used (Fig. 8*d*). This structure consists of pig insulin (Thr B30 → Ala), which potentially could induce differences when compared with human insulin. Comparison of the simulated powder patterns of 4ins (Fig. 8*d*) and 1mso (Fig. 8*c*) with the experimentally obtained pattern from F crystals shows that none of them matches perfectly, but the room-temperature data do have a better overall match. For all samples, the largest intensity variations are found in the 2θ region 0.9–2.5°.

4. Discussion

In this study we have shown that powder diffraction is a valid and useful tool for the analysis of different insulin polymorphs. Even without using specialized equipment and methodology, powder patterns of the microcrystals were all characteristic for the different crystal forms and could even distinguish samples with minor structural differences. The tetragonal I and J crystals are an example of the latter: a change in the binding affinity to the basic poly arginine peptide protamine affected the diffraction pattern of the J crystals and resulted in a detectable shift in peak positions, when compared with the I crystals. An even more pronounced difference was found between the rhombohedral crystals (D, E and F) where the different cell lengths of the c axis (*ca* 40, 37 and 34 Å, respectively; Fig. 1 and Table 1), corresponding to different conformations of the N-terminal part of the B-chain, significantly affected the powder pattern. The major difference was found around $2\theta = 4.0°$, where maximum shifts of peak positions were observed.

PCA was shown to be useful for visualization and comparison of multiple samples. The visual information presented in the two- and three-dimensional PCA score plots is easier to interpret than the two-dimensional intensity *versus* 2θ plots. The PCA was loaded with the full profile data in the 2θ range 0.9° to 6°. There is a clear benefit from using the full profile data compared with discrete peak position matching: the full profile is more forgiving of small shifts in peak position. In addition, no subjective peak extractions or tolerance cut-offs need to be applied by the user. Nonetheless, the analysis method is still sensitive to both peak position and peak intensity, which was important since it is directly related to changes in cell dimensions and unit-cell content.

Although indexing of the samples would take the analysis a step further, our attempts to index the samples using the programs *DICVOL* (Boultif & Louer, 2004), *TREOR* (Werner *et al.*, 1985) and *ITO* (Visser, 1969) did not succeed. Rescaling of the d-spacing by dividing the wavelength by a certain factor is a common approach used for indexing of protein powder data, but this was not successful. The reason is probably that the peak resolution is too low, resulting in peak overlap. It is well known that powder data collected on area detectors display poorer resolution compared with more specialized powder diffraction setups. Our medium-resolution powder diffraction profiles are not sufficiently resolved for successful indexing, but are still useful for effective classification of the crystal system.

The applicability of the method to larger proteins than insulin has not been tested yet, but should be possible even in cases where the unit-cell dimensions are larger. The orthorhombic insulin crystals used in this study have quite large cell axes and three hexamers in the asymmetric unit. The powder pattern has its major peaks in the low-2θ region, but a characteristic pattern still results.

The powder patterns calculated from atomic coordinates match the observed patterns well enough to identify the crystal system. No explicit Lorentz or polarization correction was applied to the observed powder diffraction data. These measured intensities are thus affected by a Lorentz factor of $1/\sin\theta$ (Warren, 1990). In order to bring the calculated intensities (based on coordinates) onto the same scale, they were multiplied by the same factor. Even after applying this factor, there are a number of differences in the intensity distribution which can be attributed to either systematic errors in data collection and processing, or to the use of an incomplete or not completely correct atomic model for calculation of the powder pattern, as follows.

(i) Although the capillaries were rotated during data collection in order to reduce the influence of preferred crystal orientations, we cannot rule out that some of our more needle-shaped crystals are oriented along the capillary, thus skewing the intensities.

(ii) The most intense and best determined peaks in the powder pattern are in the low-angle region. In contrast, protein structures from single-crystal data are refined towards agreement of high-angle reflections. Often a low-resolution cut-off is used. The measured intensities of low-angle reflections are strongly affected by bulk solvent. Inclusion of an appropriate bulk-solvent model has in other cases been shown to improve the agreement of calculated and observed powder patterns (Von Dreele, 2005).

(iii) Differences could also originate from small structural differences between the larger crystals used for single-crystal structure analysis and the microcrystals. Larger crystals are often grown under slightly different crystallization conditions, which might induce structural changes in the protein.

(iv) In the case of the tetragonal I and J crystals, there is no solved structure with a detailed description of the protamine binding. The protamine is therefore not included in the PDB files, and thus cannot be accounted for in the calculated powder pattern.

(v) Structural differences could also be induced by cryo-cooling. All powder data were collected at room temperature, while some of the single-crystal structures used here were determined at 100 K, which in some cases alters the cell constants and can induce structural differences (Smith *et al.*, 2003; Halle, 2004). Both the D and F crystals exhibited a small shift between experimental and predicted patterns above $2\theta \simeq 4°$. Comparing the F crystals with the room-temperature pig insulin structure 4ins (Fig. 8*d*) improved the agreement. Remaining differences are probably due to different amino acid sequences or actual differences in cell parameters between single crystals and powder crystals. Likewise, the observed differences between the experimental and simulated

D pattern are most likely also an effect of cryo-cooling-induced changes of the cell parameters.

We conclude that the use of medium-resolution X-ray powder diffraction is a valuable tool for characterization and evaluation of microcrystal suspensions of proteins, both during new formulation and polymorph screenings, and in manufacturing process control. An example of the usefulness is illustrated with the unknown insulin crystals X. This microcrystalline suspension was produced with an in-house formulation screen. The powder diffraction pattern is clearly different from other known crystal forms. Ongoing studies currently aim at identifying and further characterizing this formulation. The identification of a formulation with novel crystallographic properties has encouraged us to use powder diffraction routinely as a tool in daily research. It has also been important for identification and verification of batch-to-batch deviations during large-scale crystallization in the production process. It should be noted that although in this study we only used the intensities as a function of 2θ values as variables in the PCA score plots, it should be possible to include other types of information in the data analysis. For example, a combination of powder data with crystallization conditions in the PCA should make it possible to study the influence of various additives and various parameters like ion strength, pH, *etc.*, in the crystallization media. It should also be possible to use powder diffraction routinely as a tool to verify crystallization screens by discriminating microcrystals from amorphous precipitate.

The authors would like to thank Lene Drube for technical assistance, Charlotte Hammelev for providing the production samples, Per Balschmidt, Helle Birk Olsen and Niels C. Kaarsholm for fruitful discussions and valuable input, and the beamline staff at the Max-lab synchrotron. The work was supported by the VTU (Ministry of Science, Technology and Innovation), Denmark, and the Novo Nordisk CORA Training and Research Program.

References

Baker, E. N., Blundell, T. L., Cutfield, J. F., Cutfield, S. M., Dodson, E. J., Dodson, G. G., Hodgkin, D. M., Hubbard, R. E., Isaacs, N. W. & Reynolds, C. D. (1988). *Philos. Trans. R. Soc. London Ser. B*, **319**, 369–456.

Balschmidt, P., Hansen, F. B., Dodson, E. J., Dodson, G. G. & Korber, F. (1991). *Acta Cryst.* B**47**, 975–986.

Balschmidt, P. (1996). AspB28 insulin crystals, Novo Nordisk A/S, US Patent US5547930.

Banting, F. G. & Best, C. H. (1922). *J. Lab. Clin. Med.* **7**, 251–266.

Barr, G., Dong, W. & Gilmore, C. J. (2004*a*). *J. Appl. Cryst.* **37**, 658–664.

Barr, G., Gilmore, C. J. & Paisley, J. (2004*b*). *J. Appl. Cryst.* **37**, 665–668.

Basso, S., Fitch, A. N., Fox, G. C., Margiolaki, I. & Wright, J. P. (2005). *Acta Cryst.* D**61**, 1612–1625.

Boultif, A. & Louer, D. (2004). *J. Appl. Cryst.* **37**, 724–731.

Brange, J. (1987). *Galenics of Insulin*. Berlin: Springer-Verlag.

Cerenius, Y., Stahl, K., Svensson, L. A., Ursby, T., Oskarsson, A., Albertsson, J. & Liljas, A. (2000). *J. Synchrotron Rad.* **7**, 203–208.

Ciszak, E. & Smith, G. D. (1994). *Biochemistry*, **33**, 1512–1517.

Ciszak, E., Beals, J. M., Frank, B. H., Baker, J. C., Carter, N. D. & Smith, G. D. (1995). *Structure*, **3**, 615–622.

Derewenda, U., Derewenda, Z., Dodson, E. J., Dodson, G. G., Reynolds, C. D., Smith, G. D., Sparks, C. & Swenson, D. (1989). *Nature (London)*, **338**, 594–596.

Gursky, O., Badger, J., Li, Y. & Caspar, D. L. (1992). *Biophys. J.* **63**, 1210–1220.

Hallas-Møller, K. (1956). *Diabetes*, **5**, 7–14.

Hallas-Møller, K., Petersen, K. & Schlichtkrull, J. (1951). *Ugeskr Laeger.* **113**, 1761–1767.

Halle, B. (2004). *PNAS*, **101**, 4793–4798.

Harding, M. M., Hodgkin, D. C., Kennedy, A. F., O'Conor, A. & Weitzmann, P. D. (1966). *J. Mol. Biol.* **16**, 212–226.

Kaarsholm, N. C., Ko, H.-C. & Dunn, M. F. (1989). *Biochemistry*, **28**, 4427–4435.

Krayenbuhl, C. & Rosenberg, T. (1946). *Rep. Steno. Mem. Hosp. Nord. Insulinlab.* **1**, 60–73.

Mammen, C. B., Ursby, T., Cerenius, Y., Thunnissen, M., Als-Nielsen, J., Larsen, S. & Liljas, A. (2002). *Acta Phys. Pol. A*, **101**, 595–602.

Margiolaki, I., Wright, J. P., Fitch, A. N., Fox, G. C. & Von Dreele, R. B. (2005). *Acta Cryst.* D**61**, 423–432.

Pechenov, S., Shenoy, B., Yang, M. X., Basu, S. K. & Margolin, A. L. (2004). *J. Control. Release*, **96**, 149–158.

Richards, J. P., Stickelmeyer, M. P., Frank, B. H., Pye, S., Barbeau, M., Radziuk, J., Smith, G. D. & DeFelippis, M. R. (1999). *J. Pharm. Sci.* **88**, 861–867.

Schlichtkrull, J. (1958). *Chemical and biological studies on insulin crystals and insulin zinc suspensions* (thesis). Copenhagen: Ejnar Munksgaard.

Smith, G. D. & Ciszak, E. (1994). *Proc. Natl. Acad. Sci.* **91**, 8851–8855.

Smith, G. D. & Dodson, G. G. (1992). *Proteins*, **14**, 401–408.

Smith, G. D. & Blessing, R. H. (2003). *Acta Cryst.* D**59**, 1384–1394.

Smith, G. D., Ciszak, E., Magrum, L. A., Pangborn, W. A. & Blessing, R. H. (2000). *Acta Cryst.* D**56**, 1541–1548.

Smith, G. D., Pangborn, W. A. & Blessing, R. H. (2003). *Acta Cryst.* D**59**, 474–482.

Visser, J. (1969). *J. Appl. Cryst.* **2**, 89–95.

Von Dreele, R. B. (1999). *J. Appl. Cryst.* **32**, 1084–1089.

Von Dreele, R. B. (2001). *Acta Cryst.* D**57**, 1836–1842.

Von Dreele, R. B. (2005). *Acta Cryst.* D**61**, 22–32.

Von Dreele, R. B., Stephens, P. W., Smith, G. D. & Blessing, R. H. (2000). *Acta Cryst.* D**56**, 1549–1553.

Warren, B. E. (1990). *X-ray Diffraction.* New York: Dover.

Werner, P. E., Eriksson, L. & Westdahl, M. (1985). *J. Appl. Cryst.* **18**, 367–370.

Whittingham, J. L., Chaudhuri, S., Dodson, E. J., Moody, P. C. E. & Dodson, G. (1995). *Biochemistry*, **34**, 15553–15563.

Wold, S., Esbensen, K. & Geladi, P. (1987). *Chemom. Intell. Lab. Syst.* **2**, 37–52.