SLAC-PUB-9385
August 2002

Automatic Alignment of X-ray Beams

Zachary R. Anderson
ERULF Program
Carnegie Mellon University
Stanford Linear Accelerator Center
Menlo Park, California

16 August 2002

Prepared in partial fulfillment of the requirements of the DOE Energy Research Undergraduate Laboratory Fellowship under the direction of Ana Gonzalez in the Macromolecular Crystallography Group at the Stanford Linear Accelerator Center.

Work supported in part by the Department of Energy contract DE-AC03-76SF00515.

Automatic Alignment of X-ray Beams. ZACHARY R. ANDERSON (Camegie Mellon University, Pittsburgh, Pennsylvania, 15213) ANA GONZALEZ (Stanford Linear Accelerator Center, Menlo Park, California, 94025).

Protein crystals and other biological samples diffract weakly in X-rays. It is therefore important that the X-ray beam be very stable. Cubic smoothing splines were fit to ion chamber counts versus the vertical position of the sample. The extrema, inflection points about the maximum, and other information about the spline were calculated to determine whether the data corresponded to a beam profile. The algorithm developed here correctly identified the absence of a beam profile in all data gathered over the course of a year from four beam lines at the Stanford Synchrotron Radiation Laboratory. This algorithm is effective and can be adapted to other beam lines.

Table of Contents

Introduction

Crystallography has become important in the study of biology. X-ray crystallography is used to determine the three-dimensional structure of large proteins and other macromolecules (Mathews, 1997). Important structural determinations range from the photosynthetic reaction center (Deisenhofer et. al., 1985) of plants to various viral structures (Rossmann and Johnson, 1989). In addition, X-ray analysis has been used to engineer proteins of greater stability (Matsumura et. al., 1989) than those normally produced by living organisms. The results of X-ray crystallography will also lead to anti-influenza drugs (Colman, 1994) and to drugs that can inhibit the HIV protease (Ridky and Leis, 1995). Crystallography is an effective tool that will continue to provide insight into many biological structures and processes.

It is important that the X-ray beam has certain properties if an experiment is to be successful. Protein crystals diffract weakly if a small volume is exposed to the X-ray beam, or if an improper wavelength is chosen. In addition, it is advantageous if the beam can be finely tuned to account for the size, temperature, and other properties of the sample (Helliwell, 1997). It is therefore important that the sample be placed in the most intense part of the beam profile so that the user is certain the properties of the beam are correct. Furthermore, small fluctuations in the electron beam patch, or heat-induced distortions of the optical elements, can cause the position of the beam to shift slightly over the course of an experiment. The position of the sample must be adjusted periodically to account for the changes in the X-ray beam.

The X-ray beam intensity is determined by passing the beam through a helium ion chamber and counting the number of atoms ionized by the beam. Software previously in use at the Stanford Synchrotron Radiation Laboratory uses a simple algorithm that takes ion chamber readings over a range of table positions and moves the sample to the position of the highest reading. This method is prone to finding false maximums due to spurious high detector readings, or from scanning the beam far from its actual maximum. In addition, the software fails to provide the user with information about any problems there may be when a scan does not contain a beam profile. Since a major goal in the field of X-ray crystallography is to automate and accelerate data collection (Mathews, 1997), this is a process that needs to be improved.

We developed and implemented a more sophisticated and statistically sound algorithm that is both reliable and efficient. Cubic smoothing splines (Green and Silverman, 1994; Reinsch, 1967) were fit to the data using a fast procedure for calculating the minimum cross validation (Hutchinson, 1986). Information about the splines was calculated and used to make a decision about the quality of the scan and the existence of a beam profile. These functions were integrated into existing data collection software. With software that is more reliable, experimenters will be able to collect more data with greater efficiency.

Algorithm Description

Scans of ion chamber counts give noisy data. There is no obvious parametric model that can succeed in fitting the many different types of beam profiles that are observed. For these reasons, what we would like is a non-parametric model that makes some trade-off between the closeness of the fit, and the smoothness of the fit.

Let the sequence of numbers $t_1,\ldots,t_n$ on an interval [a,b] be given such that the sequence is strictly increasing. The function g defined on [a,b] is a cubic spline if two conditions are met.

First, g must be a cubic polynomial on each interval $(a,t_1),(t_1,t_2),\ldots,(t_n,b)$. Second, each of the polynomials must fit together such that g and its first and second derivatives are continuous at each t_i. A cubic spline is uniquely identified by its value and second derivative at each t_i . See Green and Silverman for a proof of this.

The procedure CUBGCV written by Hutchinson in FORTRAN implements the linear time algorithm by Reinsch for fitting a cubic smoothing spline to n noisy data points where the spline function g minimizes the penalized sum of squared residuals:

$$S(g) = \Sigma\{Y_i - g(t_i)\}^2 + \alpha\int\{g``(x)\}^2 dx \qquad (1).$$

The integral of the squared second derivative of g is a measure of the 'smoothness' of g. As α approaches zero the spline becomes an interpolating spline. As α approaches infinity the spline becomes a linear regression. Penalizing the sum of squared residuals is how the trade-off between a close fit and smoothness is made.

Reinsch's algorithm finds the values of g and its second derivative at each t_i for a particular value of α. If **g** is the vector of spline values at each t_i, and **Y** is the vector of data values at each t_i then:

$$\mathbf{g} = \mathbf{Y} - \alpha Q\gamma, \qquad (2)$$

where γ is the vector of second derivatives at each t_i, α is a smoothing parameter, and Q is defined as an n by n-2 matrix with entries q_{ij} for i=1,…,n and j=2,…,n-1, given by:

$$q_{j-1,j} = h_{j-1}^{-1},\ q_{jj}=-h_{j-1}^{-1}-h_j^{-1}, \text{ and } q_{j+1,j}=h_j^{-1} \qquad (3)$$

with $h_i=t_{i+1}-t_i$, and $q_{ij} = 0$ for $|i-j| \geq 2$. The vector of second derivatives, γ, is given by:

$$(R+\alpha Q^TQ)\gamma = Q^T\mathbf{Y} \qquad (4)$$

where R is an n-2 by n-2 matrix with elements r_{ij} given with i and j running from 2 to n-2, given by:

$$q_{ii} =(h_{i-1} + h_i)/3 \text{ for } i = 2,\ldots,n-1, \quad (5)$$

$$r_{i,i+1} = r_{i+1,i} = h_i/6 \text{ for } i = 1,\ldots,n-2, \quad (6)$$

and $r_{ij} = 0$ for $|i-j| \geq 2$. The equation for γ can be solved using Cholesky decomposition in linear time. The vectors **g** and γ then describe cubic polynomials between each two t_i. It is now straightforward how to calculate the value of the spline at every point in the domain.

Hutchinson also includes a linear time algorithm for minimizing the generalized cross validation (GCV) which gives an α that minimizes S(g):

$$GCV(\alpha) =S(g)/tr\{I-A(\alpha)\}^2, \quad (7).$$

where I is the identity matrix, and $A(\alpha)$ is the hat matrix:

$$A(\alpha) = (I + \alpha QR^{-1}Q^T)^{-1}\mathbf{Y}, \quad (8)$$

These algorithms were used to fit a cubic smoothing spline to ion-chamber counts versus the vertical position of the sample. The maximum and minimum of the spline function were found analytically along with the inflection points about the maximum if they existed. Figure 1 is an example of data and the fitted spline that corresponds to a beam profile. We note the existence of a well defined peak and maximum.

The algorithm developed here rejects scans not including an entire X-ray beam profile based on the following criteria. First, any scan containing less than a third of non-zero counts is rejected.

These kinds of scans are rejected because they do not contain enough information to justify moving the sample. Figure 2 is an example of this type of scan.

Any scan where the maximum lies on the boundary of the domain of the spline is rejected. These types of scans are rejected because they do not contain an entire profile of the beam. Figure 3 is an example of this type of scan.

Scans where one of the inflection points about the maximum lies on the boundary of the domain are rejected. These scans are rejected because the maximum occurs too close to the boundary. The second derivative is always exactly zero at the boundary of the domain of a natural cubic smoothing spline. Therefore, no useful information about the beam profile is gained by identifying these points as good inflection points. Figure 4 is an example of this type of scan.

Any scan where the ratio of the maximum to the minimum as scaled by the flux of the particular beam line is too small is rejected. This criterion allows scans where the maximum is not significantly different from the minimum to be rejected. A lower bound on this ratio can be provided to the algorithm as a parameter. Figure 5 is an example of this type of scan.

Any scan where the distance between inflection points is too small is also rejected. This allows peaks which are too narrow to be representative of the beam profile to be rejected. A lower bound on the distance between inflection points may be provided to the algorithm as a parameter. Figure 6 is an example of this type of scan.

Finally, scans where the magnitudes of the gradients at the inflection points as scaled by the maximum of the scan are too large are rejected. This allows scans which are too steep to be representative of the beam profile to be rejected. Upper bounds on the magnitudes can be provided to the algorithm as parameters. Figure 7 is an example of this type of scan.

Results

The algorithm was able to correctly identify all data not containing a beam profile from all scans performed between November 27, 2001 and July 2, 2002 on four beam lines at the Stanford Synchrotron Radiation Laboratory. There were approximately five thousand scans in this period. The algorithm failed to correctly identify the existence of a beam profile in only one scan from this period (Figure 8). Therefore, the algorithm gave no false positives and only one false negative over all the data.

We were able to arrive at values of the parameters described above for the four beam lines. These parameters were carefully chosen by examining the properties of scans known to contain beam profiles. Tables 1 through Table 4 show details of important values from all scans with beam profiles from the four beam lines.

Discussion

The algorithm performs very well on the data available. The absence of any false positives is clearly very promising. In the development of the algorithm, four parameters were added with the sole motivation of eliminating a small number of false positives. This is noteworthy because it can be concluded that the algorithm performs reasonably well even without the introduction of parameters. To claim that the parameters are unnecessary is probably shortsighted, however. The parameters add needed flexibility to the algorithm. Every beam line is different, and properties of the individual beam lines can be altered. Making adjustments to beam attenuation and energy will almost certainly change what a good beam profile looks like. Since for practical purposes false positives are unacceptable, it is important to be able to finely tune the algorithm by using the parameters.

Table 5 lists the values of parameters for all beam lines that gave the above results. The minimum width of the peak for each beam line corresponds to typical widths set by users. Setting the minimum value of the ratio between the maximum and the minimum scaled by the flux to larger than a tenth ensures that the maximum is effectively one order of magnitude larger than the minimum. The gradients at the inflection points around the maximum are scaled by the maximum because it is reasonable to expect large gradients if the maximum is large and small gradients if the maximum is small.

It is important for the usefulness of this algorithm that the parameters be adjusted carefully to match any changes to a beam line. Great care has been taken to ensure that this algorithm can be applied in general to similar problems. However, it is essential that enough data from scans of the beam profile be available so that an informed decision can be made about how to set the parameters.

The information that is gathered by this algorithm is helpful in two ways. First, It can be used to automatically move the sample to the optimally aligned position. Second, if the data does not correspond to the beam profile, the user is notified and given several hints about what could have gone wrong. This reaches the goal of improving automation of data collection.

More detailed information about the algorithm including C source code can be found in the BLU-ICE data collection software in use at the Stanford Synchrotron Radiation Laboratory and several other synchrotron light sources around the world. BLU-ICE is an Open Source software project and is used and developed by people in several countries.

Acknowledgements

Portions of this research were carried out at the Stanford Synchrotron Radiation Laboratory, a national user facility operated by Stanford University on behalf of the U.S. Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research, and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program, and the National Institute of General Medical Sciences.

I thank the Department of Energy, and the Stanford Linear Accelerator Center (SLAC) for creating, funding, and organizing the Energy Research Undergraduate Laboratory Fellowship (ERULF).

I thank the Macromolecular Crystallography group, especially my mentor Ana Gonzalez, and Henry van den Bedem for their support, encouragement, and expertise. I also thank Sekazi Mtingwa and Helen Quinn for their hard work in organizing the ERULF program at SLAC.

Bibliography

Colman P. 1994. Structure-based drug design. Curr. Opin. Struct. Biol. 4, 868.

Deisenhofer, J. et al. 1985. Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3 Angstroms resolution. Nature 318, 618.

Green, P., Silverman B. 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman and Hall, NewYork.

Helliwell, J. R. 1997. Overview of Synchrotron Radiation and Macromolecular Crystallography. Methods in Enzymology 276. Academic Press, New York.

Hutchinson, M. 1986. A Fast Procedure for Calculating Minimum Cross Validation Cubic Smoothing Splines. ACM TOMS 12: 150.

Matsumura M. et. al. 1989. Nature 342, 291.

Matthewes, B.W. 1997. Recent Transformations in Structural Biology. Methods in Enzymology 276. Academic Press, New York.

Reinsch, C.H. 1967. Smoothing by Spline Functions. Numerical Mathematics 10:177.

Ridky, T., Leis J. 1995. Biol. Chem. 270, 29621.

Rossman, M., Johnson, J., 1989. Ann. Rev. Biochem. 58, 533.

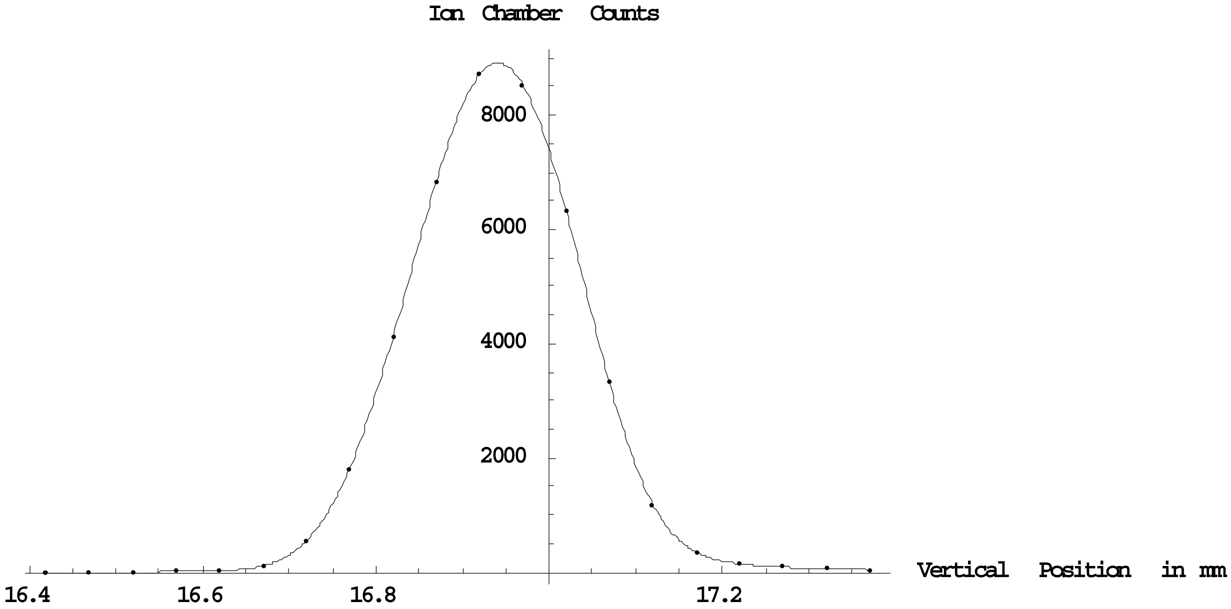

Figure 1: Ion chamber counts vs. Vertical Sample Position from beamline 11-1 at the Stanford Synchrotron Radiation Laboratory. This is a good beam profile.

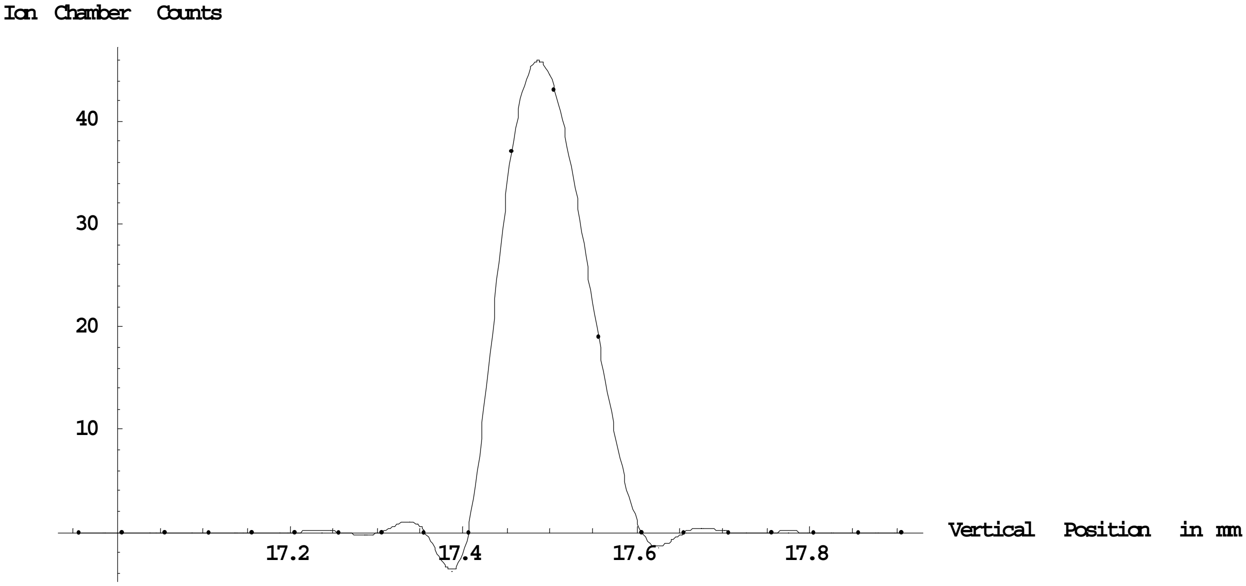

Figure 2: Scan from beamline 11-1. All but three of the counts are non-zero. This is not a good profile of the beam.

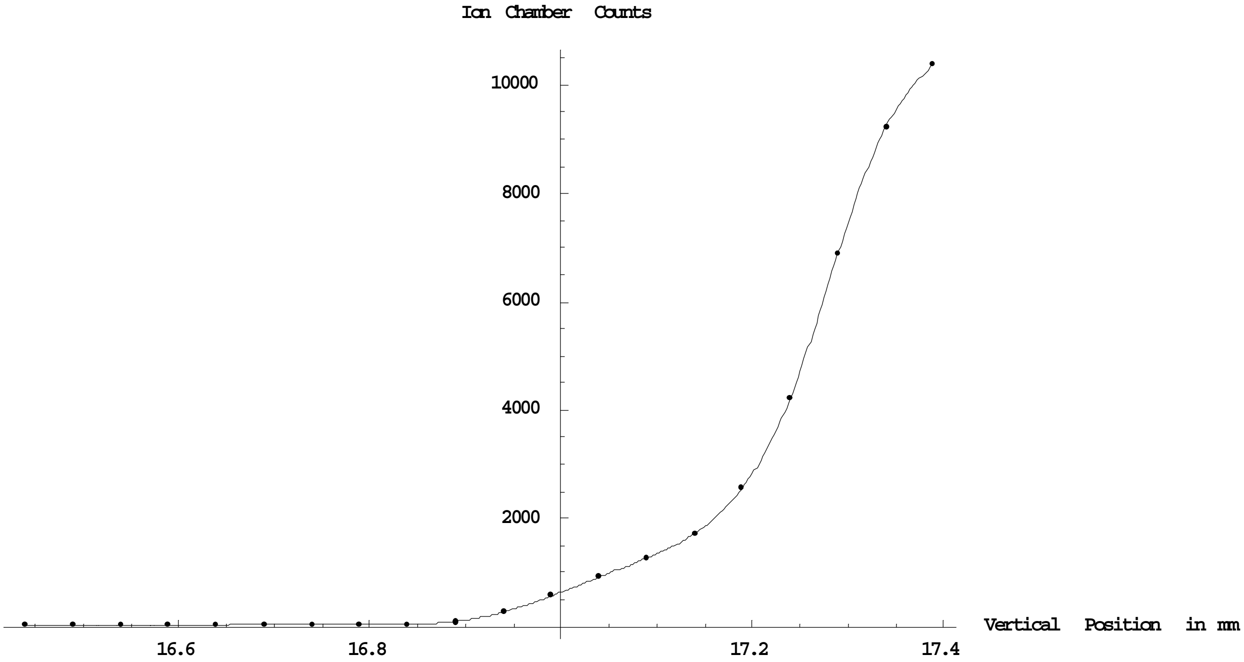

Figure 3: Scan from beamline 11-1. The maximum lies on the boundary of the domain. This is not a good profile of the beam.

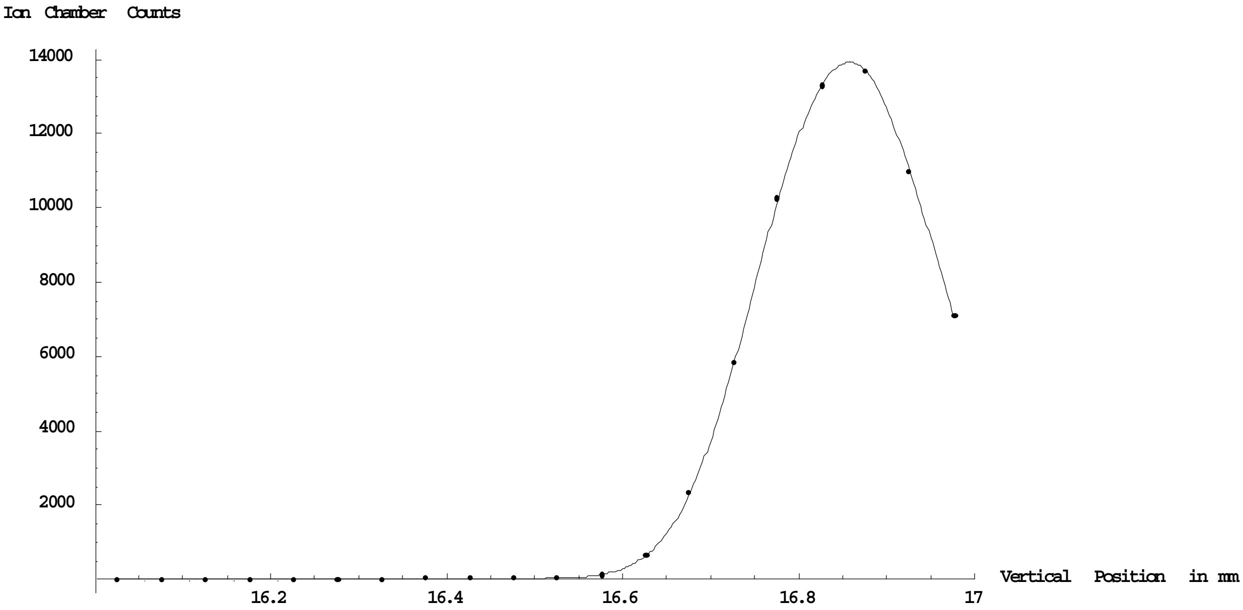

Figure 4: Scan from beamline 11-1. The right inflection point is on the right boundary of the domain. The scan does not contain an entire beam profile and must be rejected.

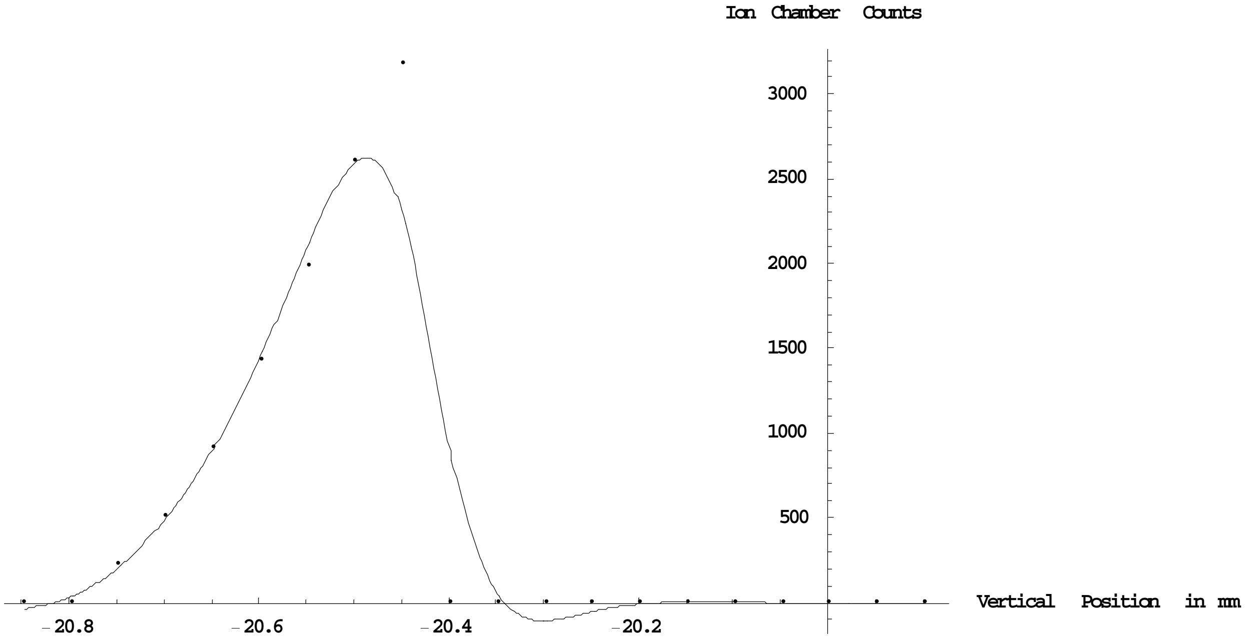

Figure 5: Scan from beamline 11-1. The ratio of the maximum to the minimum is too small for this scan to correspond to a beam profile.

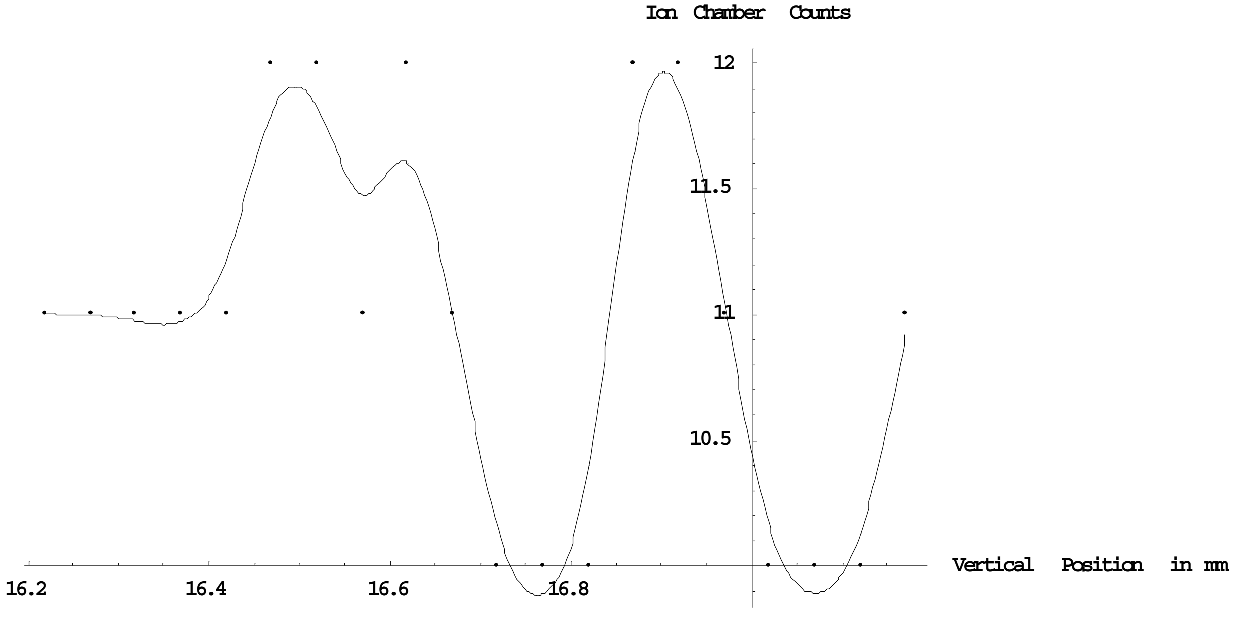

Figure 6: Scan from beamline 9-2. The counts reach their maximum value and then drop immediately to zero. The scan is rejected because the peak is too narrow.

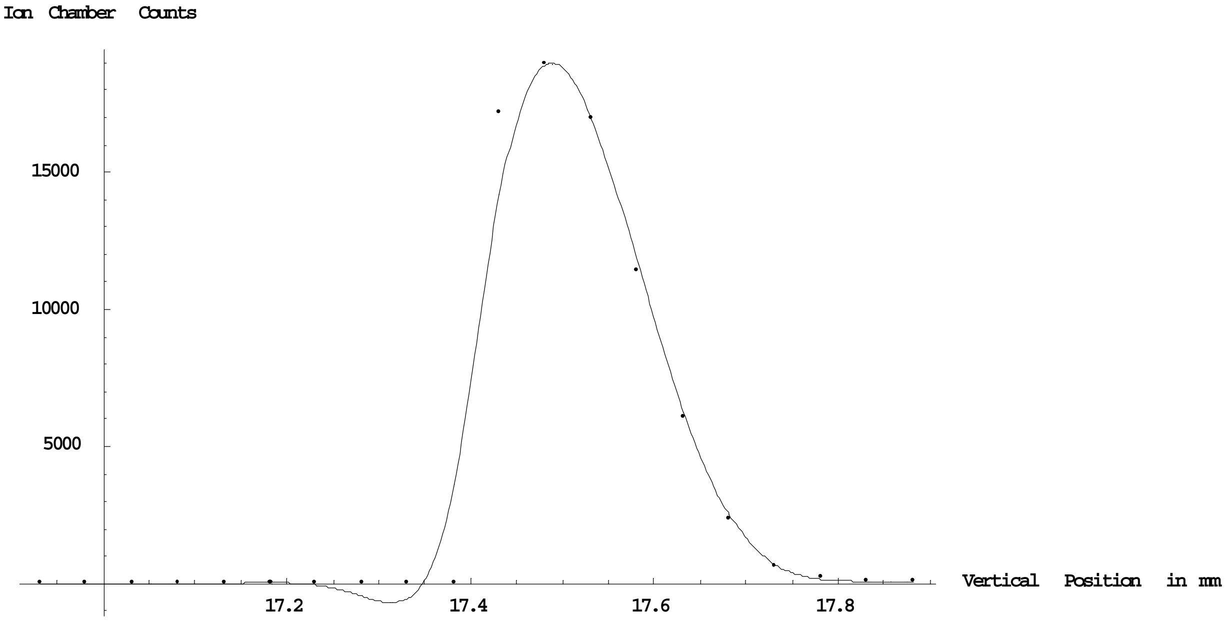

Figure 7: Scan from beamline 11-1. This scan is wide enough, but must be rejected because the gradient on the left side is too great.

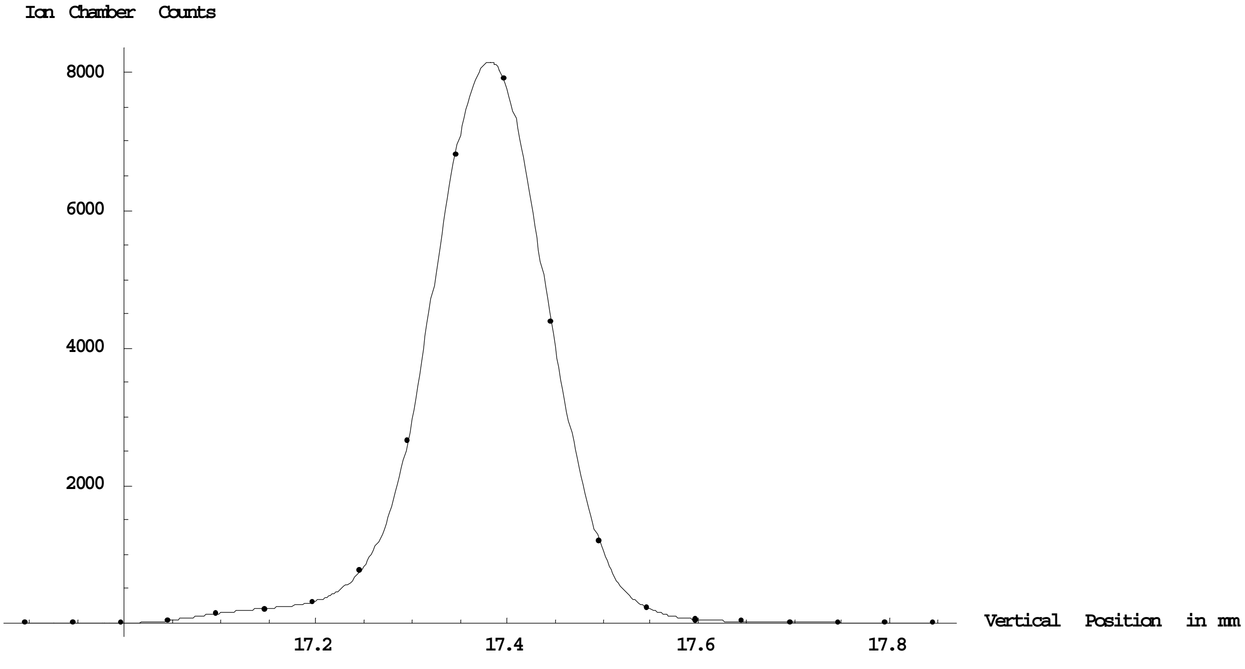

Figure 8: Scan from beamline 11-1. The scan is rejected because the gradient on the left is too large.

11-1	Maximum over all scans	Minimum over all scans	Average	Standard Deviation
Distance between inflection points	0.42084	0.10406	0.20177	0.03644
(Maximum / Minimum) / Flux	292.60602	0.22140	8.63575	17.62843
(Gradient at Left Inflection Point) / Maximum	10.28367	0.48775	5.90913	1.04081
(Gradient at Right Inflection Point) / Maximum	-1.37345	-10.91836	-6.47168	1.09003

Table 1: Data from beamline 11-1.

9-1	Maximum over all scans	Minimum over all scans	Average	Standard Deviation
Distance between inflection points	0.25223	0.18137	0.21836	0.01374
(Maximum / Minimum) / Flux	3.49800	0.38692	2.17684	0.82833
(Gradient at Left Inflection Point) / Maximum	6.31475	4.75898	5.34490	0.26656
(Gradient at Right Inflection Point) / Maximum	-5.30567	-7.20695	-5.96586	-0.29935

Table 2: Data from beamline 9-1.

9-2	Maximum over all scans	Minimum over all scans	Average	Standard Deviation
Distance between inflection points	0.53484	0.17535	0.29740	0.06027
(Maximum / Minimum) / Flux	302.83627	0.17106	41.95127	39.25404
(Gradient at Left Inflection Point) / Maximum	9.65318	0.13442	3.60694	0.61761
(Gradient at Right Inflection Point) / Maximum	-1.85311	-9.11106	-4.80498	0.79311

Table 3: Data from beamline 9-2

5-1	Maximum over all scans	Minimum over all scans	Average	Standard Deviation
Distance between inflection points	0.52106	0.15630	0.26691	0.10135
(Maximum / Minimum) / Flux	3.47548	0.71433	1.38260	0.74760
(Gradient at Left Inflection Point) / Maximum	8.78361	1.35300	3.18396	1.65024
(Gradient at Right Inflection Point) / Maximum	-0.60733	-7.01387	-3.15706	1.65830

Table 4: Data from beamline 5-1

Beam Line	11-1	9-1	9-2	1-5
Minimum Width	.1	.15	.15	.15
Minimum (Max/Min)/Flux	.15	.15	.15	.15
Maximum Gradient at Left Inflection Point	11	11	11	11
Maximum Gradient at Right Inflection Point	-12	-12	-12	-12
Flux	120	50	70	3

Table 5: Values of parameters used to eliminate all false positives and minimize false negatives.