1996

Assessment of protein crystal quality by X-ray diffraction methods.

Mark Johannes van der Woerd
University of Alabama at Birmingham

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

ASSESSMENT OF PROTEIN CRYSTAL QUALITY BY X-RAY DIFFRACTION METHODS

by

MARK JOHANNES VAN DER WOERD

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

1996

ABSTRACT OF DISSERTATION
GRADUATE SCHOOL, UNIVERSITY OF ALABAMA AT BIRMINGAM

Degree Ph.D. Program Chemistry

Name of Candidate Mark Johannes van der Woerd

Committee Chair Dr. Lawrence J. DeLucas

Title Assessment of Protein Crystal Quality by X-ray Diffraction Methods

Protein crystal growth is notoriously difficult and its mechanisms are poorly understood. In order to improve the success rate for growing protein crystals, it is necessary to judge crystal quality. A quality descriptor will allow choice of the best crystals for structure determination. With a quality descriptor it will also become possible to link quality to growth conditions, which leads to improvement of the process of protein crystal growth, as well as general knowledge about this process. Crystal quality is traditionally judged based on size and morphology. In this work a quality descriptor is developed, based on X-ray diffraction.

X-ray data from protein crystals can be interpreted in various ways, but not all interpretations are equally useful. It is shown that both reflection intensities of individual observations in relation to their σ-value and estimates of temperature factors are objective methods for data interpretation. Both methods are independent of data set size and contents, given that the data sets are statistically sound.

Temperature factor estimates obtained by Wilson plots and $I/\sigma(I)$ curves were successfully compared for data sets obtained from crystals in three different experiments. Significant differences in data quality are shown in the comparison of hen egg white lysozyme crystals grown from standard preparations at pH 4.7, and crystals grown from

ultra-pure enzyme preparations at pH 8.2. Some of the differences observed are likely related to crystal growth conditions. The level of purification of the protein and pH are two possible causes for the differences observed. Data quality also differed significantly for lysozyme crystals grown with different temperature profiles, but it is not possible to conclude if this change in quality is caused by the crystal growth conditions, or by other experimental parameters.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

CHAPTER 1

INTRODUCTION

1.1 Problem Definition

Protein crystals are more difficult to grow than crystals which are not macromolecular in nature. The difficulties encountered are usually attributed to the nature of protein crystals, such as high solvent content of the crystal, and weak intermolecular interactions.[42] Difficulties in protein crystallization may be due to the nature of protein crystals, our relatively short experience in crystallizing proteins,[18] or our perhaps inappropriate assumption that macromolecular crystals are similar to crystals of smaller molecules. The fact is that frequently the most time consuming step in the process of solving a structure is growing suitable crystals. Once crystals are grown, there are two common criteria to judge their quality: crystal size and crystal habit. The main purpose of this work is to define a quality descriptor based on diffraction data from protein crystals. Diffraction data are of the utmost importance for structure determination. Specifically, assessing crystal quality will help structure determination because it will become possible to use only those crystals that will provide the best available data. Conversely, grading crystal quality based on diffraction data will allow for a feedback mechanism to improve crystal quality. The basis for this last statement is an assumption that there is a relationship between growth conditions and crystal quality and hence the quality of acquired data.

There are a few general conditions considered favorable by many, if not all crystallographers. The first is a high resolution limit for available data. In this work it is implicitly assumed that a data set from a protein crystal is always considered to be of better quality if the resolution limit for the data is higher. Higher resolution implies that more detail will be visible in the structure. At the same time it means that the ratio of observed data points (or reflections) and parameters to be determined, such as atomic coordinates, will become more favorable. This is important because all observations are inherently associated with errors and these errors become increasingly smaller if the system is more over-determined. Thus, if the ratio between observed reflections and atomic parameters is high, the atomic parameters will be more accurate.

The second condition generally assumed to be favorable is a high signal-to-noise ratio in the data. This parameter is not completely independent of resolution, as the signal-to-noise ratio inherently drops off with increasing resolution. A predetermined ratio at which data are considered acceptable will not be defined, because data sets, or crystals, will be compared and the question should therefore be how the signal-to-noise ratios compare in a particular resolution range. It is assumed that a high signal-to-noise ratio is preferable over a low ratio. This can be justified by considering that random experimental errors in the data become smaller when the signal-to-noise ratio increases.

It is the main focus of this work to reliably describe the changes in the product (the crystals) when growth conditions vary. This will be done by analysis of diffraction data. A short overview follows for the crystal quality descriptors found in literature.

The first and very obvious quality descriptor is crystal size. Within the experimental conditions normally accessible, large crystals are always deemed superior. This

statement is not universally true, as extremely large crystals would cause too much absorption of X-rays to be useful for structure determination. Concern about absorption problems is not warranted[5] up to sizes of 1-2 mm. This size limit is higher than the sizes of protein crystals commonly grown. One of the many parameters which influence crystal growth is gravity,[35] and a particularly significant example from the literature which uses size as a descriptor for crystal quality is that of bovine insulin.[11] The results for this experiment show that the average size of crystals grown in microgravity is increased; that this increase, through measurement of very many crystals, is statistically significant; and finally, that this result is reproducible by carrying out experiments on three different Space Shuttle flights. Thus, using crystal size as a quality assessment tool, it is shown that changing gravity conditions improved crystal quality for bovine insulin. These experiments with insulin were performed in large volumes with changes in temperature as the driving force for the crystallization. More typical are the experiments with the aid of vapor diffusion,[36] which are all performed in much smaller volumes. Only a few crystals are usually available from each experiment, and a statistical analysis as performed for bovine insulin is difficult. There are many examples in which the crystals are compared based on their size. Using the vapor diffusion method, the experiments were performed both in microgravity and under standard conditions. A few of these examples, in which size increase has been reported with crystals grown in microgravity, are: β-galactosidase,[31] lysozyme,[15,31,47] D-amino transferase,[47] satellite tobacco mosaic virus,[37] recombinant human γ-interferon D1,[10,13,14] porcine elastase,[10,13] isocitrate lyase,[10,13] lectin I from *Lathyrus ochrus*,[6,10] phospholipase A2,[10] and human serum albumin.[10,38]

The second frequently used quality descriptor is morphology. Morphology is the general shape of the crystal which, albeit somewhat subjective, refers to the perfection of the crystal habit: the sharpness of the edges, the regular shape and luster of the faces, and the lack of visible internal defects. A striking example of the change in morphology when crystals are grown in microgravity occurs for crystals of isocitrate lyase.[14] They form perfect prisms, whereas gravity conditions cause a very irregular, dendritic growth in the same crystals. Other examples have been reported for which morphology clearly changed in a reduced gravity environment, sometimes accompanied by space group change: canavalin,[10,15] human C-reactive protein,[15] lysozyme,[47] and satellite tobacco mosaic virus.[37]

A third possible quality descriptor could be the diffraction data. Because diffraction data lead to structures, it would seem a logical choice to assess crystal quality via quality of structural information. In fact, this approach was used in the case of human recombinant insulin.[44] Changing crystallization conditions from the hanging drop to the batch method, with temperature as the driving force for crystallization, it was possible to grow very high quality crystals in microgravity. These superior crystals provided access to structural information which was not available before. In the majority of structure determinations, such a comparison cannot be made. If the goal of the experiment is optimization of crystal growth conditions, structural information does not become available until much later in the process. If the goal is the exploration of protein crystal growth mechanisms, structural information will also not be available. It then becomes important to decide on two issues: which aspects of diffraction data are to be used and how much data

must be obtained for each crystal to be representative for the entire data set. These issues will be addressed in chapter 2 of this work.

Diffraction data as a crystal quality descriptor have been used infrequently, such as in the case of ribonuclease-S.[3] For this protein, plots of percentage data with $F/\sigma(F) \geq 5$ as a function of resolution are reported, as well as B-factor estimates, based on Wilson-plots. For satellite tobacco mosaic virus a plot of average I/σ as a function of $\sin^2\theta/\lambda^2$ is given.[37] For human serum albumin a plot of average $I/\sigma(I)$ as a function of $1/d^2$ is reported, very likely with a different definition of I and σ compared to the satellite tobacco mosaic virus analysis.[11] The same work shows the fraction of data with $I/\sigma(I) \geq 5$ as a function of resolution, together with relative Wilson plots, for malic enzyme crystals. Both methods used in this work are different from the data reported for ribonuclease-S. The crystals of bovine insulin were analyzed by interpretation of diffraction by means of pixel analysis for data frames.[33]

All these methods are different and it is unclear which method gives a reliable evaluation of a crystal. It is possible that more than one method provides the evaluation, or even that a combination of methods is necessary. Various methods will be evaluated in chapter 2, after formulation of criteria for the evaluation.

1.2 Background of Protein Crystal Growth

Three steps occur, in very general terms, during crystal growth. The first step is the formation of nuclei, the second the actual crystal growth, and the third is cessation of growth.[20,30] Optimal conditions, such as temperature, pH, and concentration are different for each step. For instance, the supersaturation needed for nucleation is much higher than the supersaturation needed to sustain growth.[42] This general principle is illustrated in

Figure 1, which shows two curves, one for each process: the rate of nucleation and the growth rate, both as function of supersaturation. Supersaturation is defined as the protein concentration relative to the concentration at saturation. Figure 1 is only a general illustration of principles and does not contain actual data representing any particular protein. Part of Figure 1 is available, however, for lysozyme.[48] The scales on the X- and Y-axes were chosen arbitrarily. The curves describing the nucleation and growth rates may overlap. They probably will overlap in practice, indicating that conditions exist at which both nucleation and growth can occur simultaneously. Assuming that the solubility of a protein, for instance lysozyme,[39] increases with temperature, Figure 2 can be derived from Figure 1. This new figure shows the rate of nucleation and growth as a function of temperature. As before, it is an illustration of general principles. The curves may overlap and the most important feature is that high supersaturation coincides with low temperatures and therefore the positions of the two curves have been reversed with respect to each other. Crystal growth occurs predominantly at high temperature and nucleation at low temperature. Figure 2 is only applicable under the assumption that the solubility increases with temperature, as is the case for tetragonal lysozyme.[39]

In an ideal crystal growth experiment, nuclei will be formed first. There should be a few nuclei and they should all be formed at the same time. This is necessary so that all crystals have the same time available to grow and reach their maximum dimensions. Crystal growth should be slow in order to obtain well-ordered crystals, without forming additional (secondary) nuclei. It is generally accepted that the growth rate is an extremely important parameter which influences the final crystal size and quality.[21] Following the

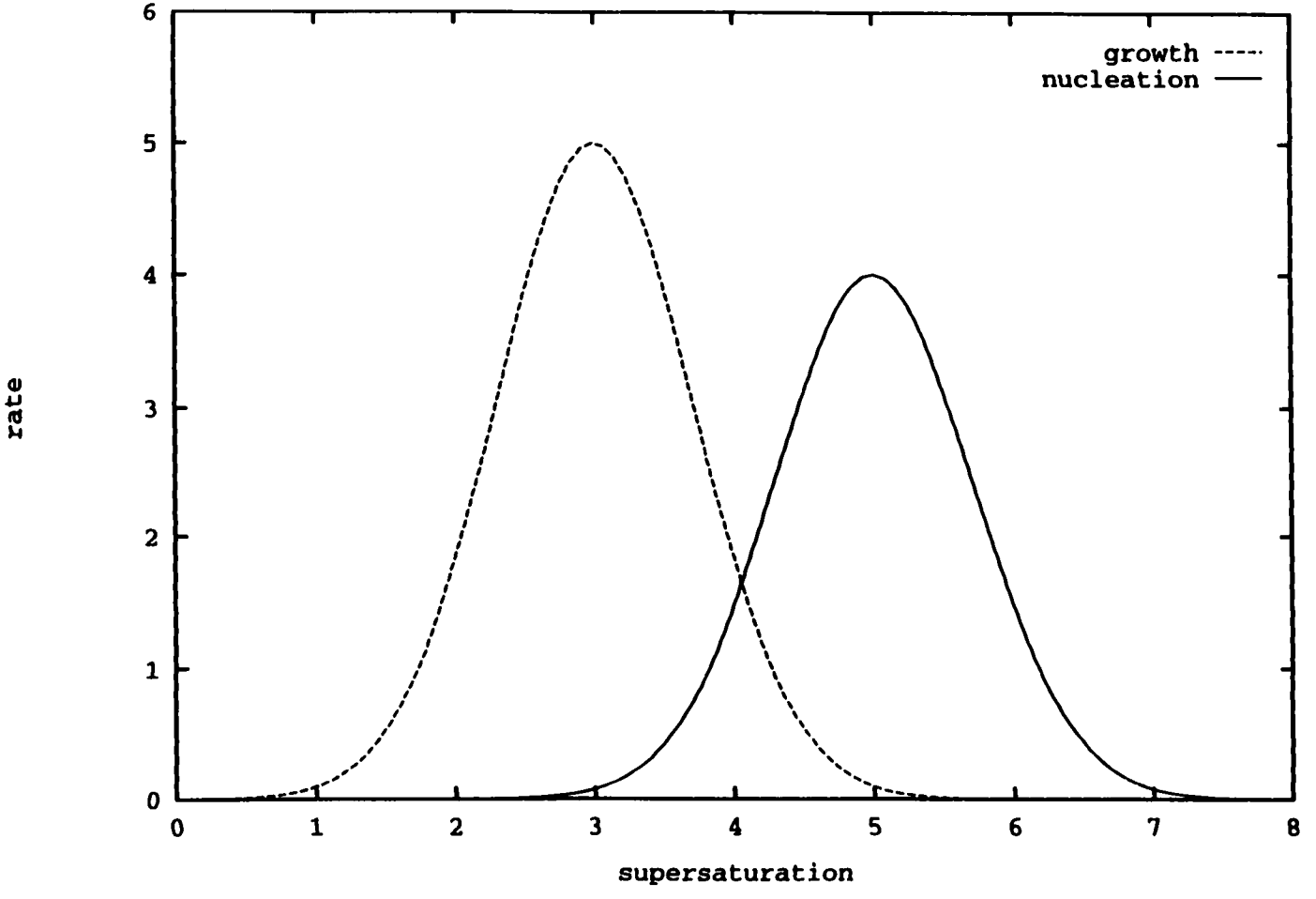

Figure 1: Plot of nucleation and growth rate as a function of supersaturation.

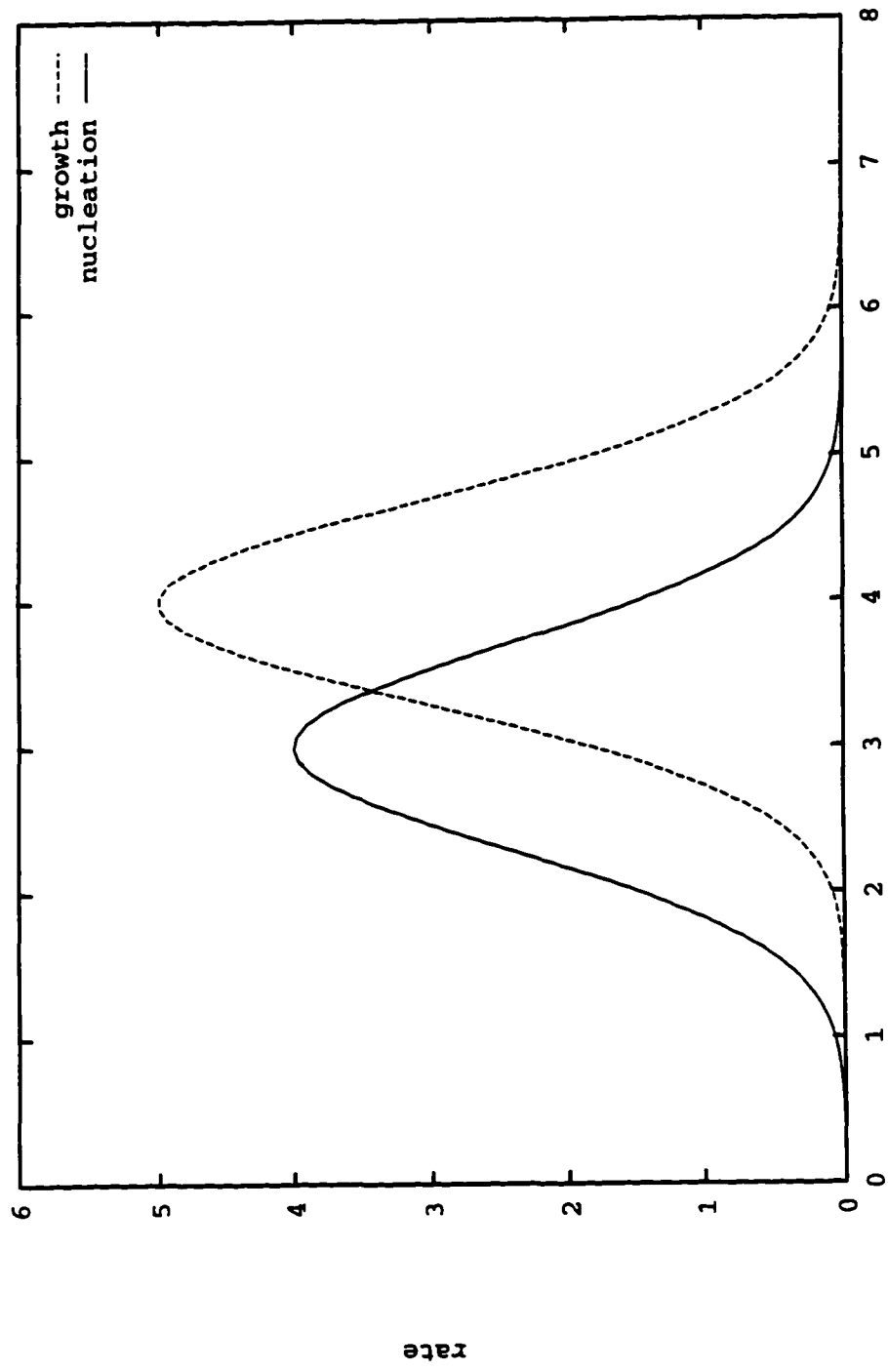

Figure 2: Plot of nucleation and growth rate as a function of temperature.

theory of inorganic (small molecule) crystal growth, a possible explanation for quality change with the growth rate, is the formation of more defects with increased growth rate.[15] Defects are caused by mismatches in the lattice, and one of the many conceivable ways a mismatch could occur is by inclusion of impurities.[34] An interesting example illustrating how macromolecular impurities influence crystal growth, possibly by creation of an imperfect lattice, is the case of lysozyme mixtures from different avian species.[1] Starting with pure lysozyme originating from one species, lysozymes with slightly different amino acid sequences were mixed in and crystal growth was monitored. Crystal size, habit, and diffraction patterns were recorded for various mixtures and mixing ratios. Thus, it was shown by intentional contamination of the protein, that crystal quality is affected by impurities.

One possible parameter to control crystal growth is temperature, since protein solubility changes with temperature. Although this is rarely the method of choice for protein crystals, temperature control gives the advantage that it is non-invasive: a closed, undisturbed system can be modified by (external) temperature regulation. Of course, data for the protein solubility as a function of temperature must be available to successfully use temperature as driving force in crystallization experiments. The use of temperature as a regulating mechanism in protein crystal growth is a continuously developing method. Evidence that the solubility indeed changes with temperature is becoming available for some proteins, like canavalin,[16] lysozyme,[8] and insulin.[4] Given that the protein solubility can be controlled by temperature, and that separate domains for nucleation and growth are accessible, it follows logically that temperature should not only be used as a driving force for crystallization, but also as a regulating tool to force either nucleation or growth

to occur. The formation of nuclei, which are small particles, considerably larger than individual protein molecules, can be detected with laser light scattering.[48] The scattering is caused by formation of nuclei, and therefore the scattering signal can be used to determine the end point of the nucleation phase and the starting point of the growth phase. Laser light scattering thus separates the nucleation and growth domains and allows for the independent optimization of the conditions in each domain. Temperature profiles can be changed by applying different slopes to the profiles, by choosing different temperatures for nucleation and growth, or by a combination of these choices. In chapter 3 an experiment in which temperature is used as the driving force in crystallization will be described as an example. The role of crystal quality assessment by means of X-ray diffraction will be discussed. The main question for this research is how the crystal growth conditions relate to the quality of the diffraction pattern, or conversely, how diffraction can be improved by changing the crystallization conditions. In the specific example in chapter 3, the crystallization conditions are changed by temperature variation.

1.3 Crystallization in Microgravity

Gravity is only one of many parameters that influence protein crystal growth.[35] It is not well understood what the influence of gravity is on protein crystal growth, or which mechanisms are at the base of the changes observed when crystals are grown in reduced gravity. Microgravity is defined as an environment in which the effective gravitational force is in the order of magnitude of 10^{-6} times the standard force on earth. One attempt to rationalize the difference between crystal growth in the earth's gravity compared to reduced gravity, is based on the density of the growth medium. Density is a function of both temperature and solute (protein) concentration. High temperature decreases the

density and high solute concentrations increase the density. The change of density and therefore local variations of concentration or temperature can be observed via the accompanying change of index of refraction.[43] Methods that allow for the measurement of these changes without disturbing the system under observation are optical in nature, such as Schlieren photography, interferometry, and holography.[43] With the aid of these techniques it is possible to show that crystal growth in the earth's gravitational field causes density differences. These differences are caused by solute depletion in the solution near the crystal and produce convective flows.[7] These flows are also occurring in the case of protein crystal growth as shown for hen egg white lysozyme,[39] but do not occur under conditions of microgravity.[10] It is proposed that microgravity may change protein crystal growth by elimination of convective flows at the crystal surface, and this explanation is referred to in many experiments.[7,12,31,37,38,40]

One study of the kinetics of lysozyme crystal growth shows the rate limiting step to be the attachment of protein molecules to the surface of the crystal. This implies a relatively fast transport of protein from the bulk solution to the surface of the crystal, presumably by convective flow.[41] The consequence of this statement is that crystal growth may be governed by completely different kinetics in microgravity. Convective flows are absent and the ratio of the protein transport rate and the attachment rate changes. The change in kinetics is used as a possible explanation for the differences observed in crystals of canavalin and satellite tabacco mosaic virus on earth versus in microgravity.[37] Observations in these experiments, such as the crystal size, optical crystal quality, and the number of crystals produced, agree with the so-called diffuso-convective model. This model is but one way of rationalizing the changes observed in

protein crystals when the growth conditions are changed. The merits of this method are still subject to discussion.[42] Other aspects of importance for crystallization in microgravity relating to microgravity are: stirring of the solution by convective flows, which causes non-uniform growth conditions providing an unequal supply of protein to the various crystal faces;[15] asymmetric growth when crystals appear on or near solid objects, such as chamber walls or other crystals;[32] and settling of crystals at the bottom of the crystallization droplet.

1.4 Temperature Factors

The use of temperature factors as a descriptor of crystal quality will be discussed in chapter 2. Although estimates for temperature factors are used in every *ab initio* protein structure determination, the information about their background, method of estimation, and their limitations is widely scattered throughout the literature. In the following sections the background as well as a method for temperature factor determination will be reviewed.

1.4.1 The background of temperature factors. X-rays, which are a form of electromagnetic radiation, interact with electrons associated with atoms. The atomic scattering factor f describes the coherent scattering of X-rays by an atom. Equations 1 through 3 are a summary of a section on atomic scattering factors from the literature.[29] The scattering factor f relates to the electron wave function ψ as described in equation 1.

$$f(\mathbf{s}) = \int \psi^{*}(\mathbf{r}) \exp(i\mathbf{s} \cdot \mathbf{r}) \psi(\mathbf{r}) dv \qquad (1)$$

The electron density $\rho(\mathbf{r})$ is a function of the distance $\mathbf{r}$ from the center of the atom. The vector $\mathbf{s}$ bisects the angle $(180\text{-}2\theta)$ between the incident and scattered wave vectors and has a magnitude of $|\mathbf{s}| = 4\pi\lambda^{-1} \sin\theta$. This is an important point, since f

changes with sin θ/λ. In practice it is common to assume atoms to be spherically symmetric (or, to average over all directions and treat the atom as symmetric) and simplify equation 1 to

$$f(s) = \int \frac{\rho(r)\sin(sr)}{sr} dr \qquad (2)$$

where $\rho(r) = [rP(r)]^2$ and $P(r)$ is the radial part of $\psi(\mathbf{r})$. Given that the wave function ψ is known, the values for f can be calculated as function of s. These values have been tabulated[29] as well as fitted to an analytic expression in the form of equation 3, in order to make computer programming easier.

$$f(\sin\theta / \lambda) = \sum_{i=1}^{4} a_i \exp(-b_i \lambda^{-2} \sin^2\theta) + c \qquad (3)$$

An example of this function is plotted in Figure 3 with the coefficients for a carbon atom taken from literature,[29] with the label B = 0. This plot is an ideal case, for which the temperature factor is zero.

The temperature or B-factor is a measure of disorder. Atoms can be disordered in two fundamentally different ways. The first is due to thermal vibration: the atoms in a structure are not at rest, but vibrate around their equilibrium position. The higher the thermal energy of the system becomes, the stronger the vibration becomes. This vibration will cause the atoms to appear slightly displaced in a crystallographic lattice. The second is static disorder, caused by imperfection in the crystal. When entire molecules, or aggregates of molecules are displaced slightly with respect to their ideal position in the lattice, the macroscopic view of the system will render a less sharp image, with less detail, which can be expressed in a B-factor. An isotropic correction factor is used to

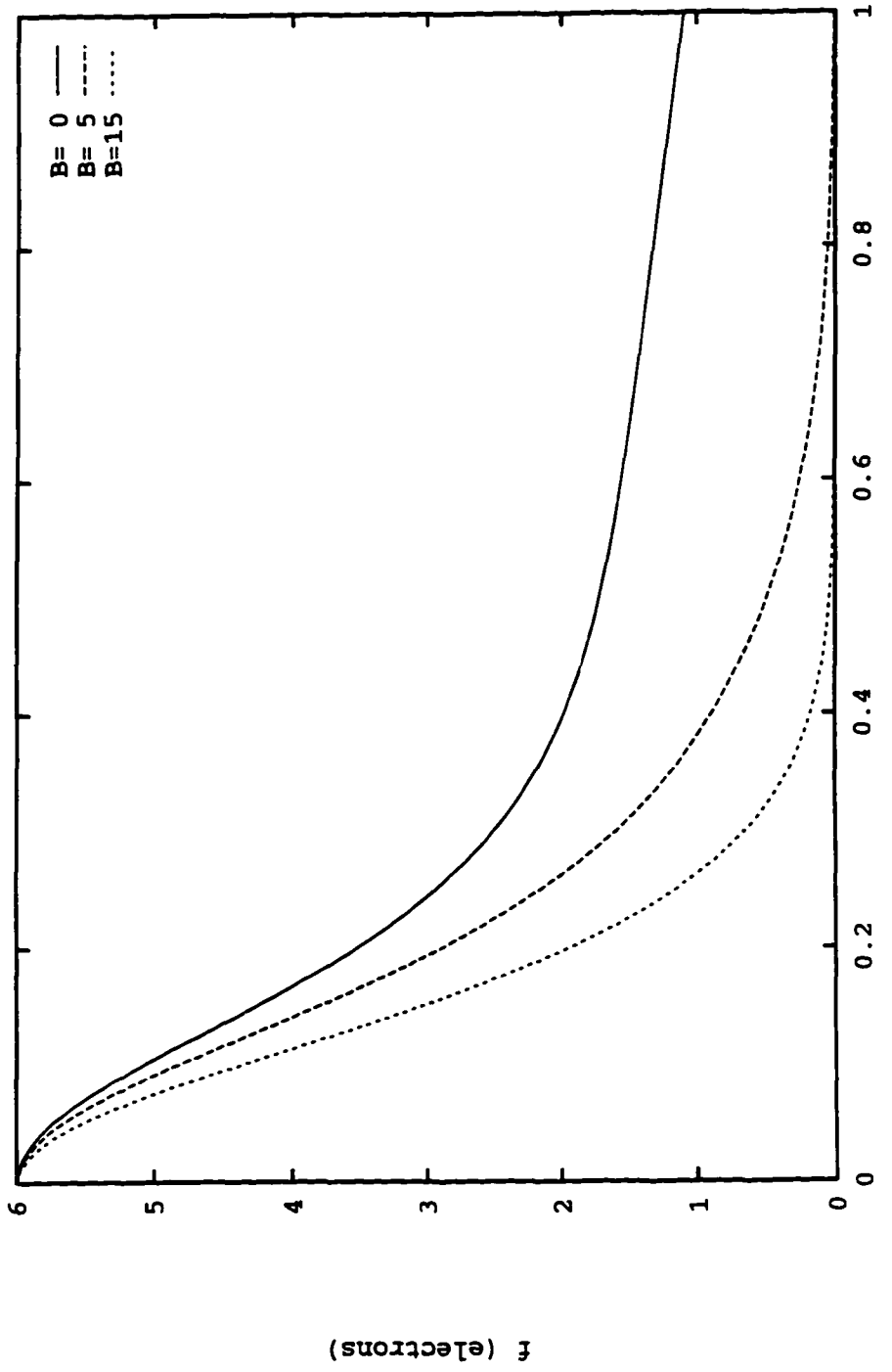

Figure 3: Plot of scattering factor *f* for carbon.

describe scattering which deviates from the ideal case due to disorder. The isotropic temperature factor B is used in this correction.

The correction factor is[17]:

$$T_{iso} = \exp\left[-B\frac{\sin^2\theta}{\lambda^2}\right] \quad (4)$$

In Figure 3 the scattering factor for carbon is plotted with B factors equal to 5 and 15, both very reasonable values for protein structures. The plot shows that larger B-factors make a dramatic difference in the scattering factor at high resolution (high $\sin\theta/\lambda$ values), whereas the change is smaller at low resolution. In the process of structure determination the scattering of large sets of atoms is measured via intensity *I*, which is proportional to $|\mathbf{F}|^2 T_{iso}^2$ and, with

$$\mathbf{F} = \sum_{j=1}^{n} f_j \exp(2\pi i \mathbf{r}_j \cdot \mathbf{s}) \quad (5)$$

When B-factors as shown in equation 4 are large, the scattering will become very weak and the reflection's intensity will become immeasurably low. The B-factors therefore limit the resolution at which a structure can be determined. Large B-factors inherently cause a lower observable resolution limit and thus detail is lost in the structure. This fact makes B-factors a prime candidate for crystal quality analysis and comparison. A B-factor is an intrinsic value, since it does not depend on the size of the crystal.

1.4.2 Theory of B-factor determination. B-factors as used in this work were first described by Wilson.[49] Originally B-factors were not determined to measure disorder, but as part of a scaling device. Wilson recognized that average intensities of reflections relate to scattering factors *f* as:

$$\overline{I_{hkl}} = \sum_a \overline{f_a^2} \quad (6)$$

This is under the assumption that the atoms are randomly dispersed in the unit cell. The procedure Wilson proposed consists of dividing the data in suitable resolution bins and calculating the ratio $\sum \overline{f}^2 / \overline{I_{hkl}}$ for each bin. The logarithm of this ratio is approximately a linear function of $(\sin\theta/\lambda)^2$. The intercept and slope of this function provide a scaling factor and a temperature factor, with the aid of which it is possible to scale two arbitrary data sets together. This method is commonly applied in isomorphous replacement. Usually relative Wilson plots are calculated, which gives the ratio of scale factors and ΔB between two data sets.[17] These two constants are then used to scale one data set to another set. It is possible to determine the scale factor and B-factor with respect to the atomic scattering factors, and this method will be used in this work.

1.5 Statement of Problem

Protein crystals are commonly judged on their size and habit. These properties, however, do not describe diffraction data, neither do they relate to structural information of the protein molecule. Since the structural information is obtained with the aid of X-ray diffraction techniques, it is necessary to develop a quality assessment tool based on diffraction. The method of quality assessment needs to be objective. This implies that a protocol needs to be developed for objective data collection, sampling, and interpretation. Various aspects of diffraction data have been used to judge data quality, but no standard is available. The aims of this work are to develop a standard technique, starting with the evaluation of currently available methods. This evaluation is necessary to ensure that the quality assessment is objective. The final recommendations of the technique to be used should include specific details for a protocol to compare diffraction data quality.

CHAPTER 2

METHODOLOGY

2.1 Overview

The goal of this work is the development of a quality assessment tool for protein crystals, based on X-ray diffraction. It is therefore necessary to look at X-ray diffraction and the data obtained with that process. The vast majority of the data sets used in this work are processed with the Xengen package.[28] The terminology used in the Xengen package will be used as a reference.

The data obtained from a diffraction experiment pass through various stages, each of which provides an opportunity to assess the quality of the data. The first stage,[28] as depicted in Figure 4, is raw data, or the data frames. These frames are collected on a Siemens multiwire area detector, and they consist of a two dimensional array of pixels. A reflection appears in the data frame as a cluster of pixels with significantly higher value than the background. The pixel values are representative of the number of photons counted in the exact location on the detector face. One approach for judging diffraction data is to look at these patterns of pixels and determine if there is a relationship between pixel pattern and data quality.

Several steps follow in the data processing path, including determination of the orientation matrix. This matrix is used to index the data. At this point the observed

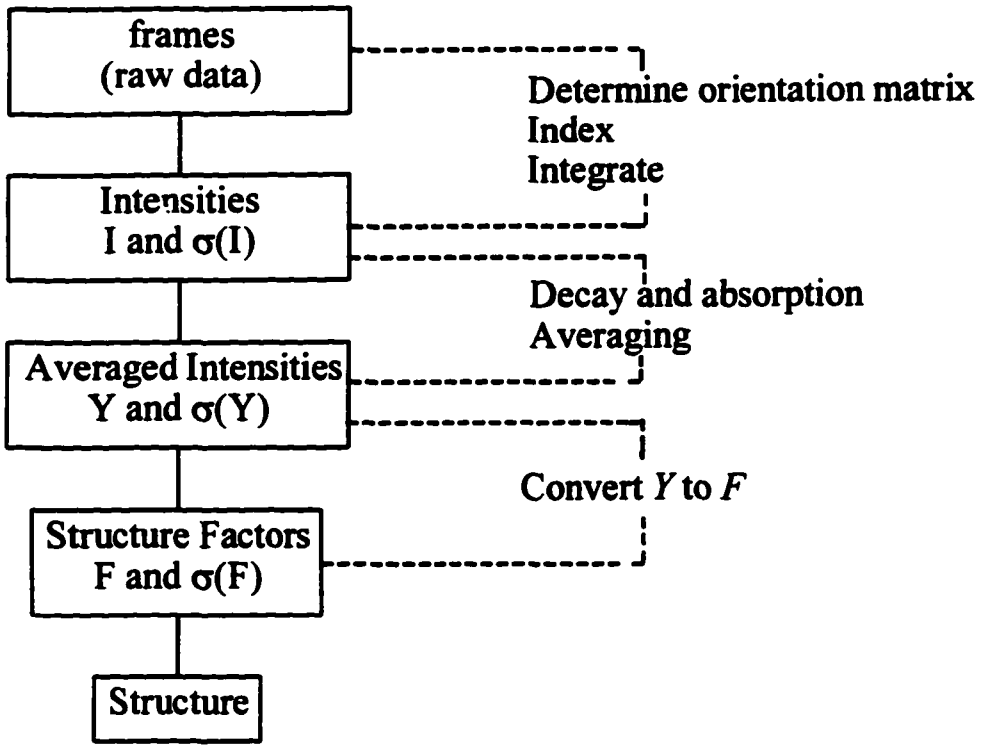

Figure 4: The various stages during data processing.

reflections are assigned their Miller indices (hkl).[46] The individual observations are integrated, which means that the values of the pixels that are considered part of the reflection are added together. This process includes determination of an estimate and removal of the background scatter. After completion of this stage, the observations have been assigned an integrated intensity value I, plus an estimate of $\sigma(I)$. Both the variance of the background scatter and the variance of the summed intensity values are included in the value of $\sigma(I)$. The intensity I has been adjusted for Lorentz and polarization corrections. The I and $\sigma(I)$ values are sometimes used as data descriptors, particularly as their ratio $I/\sigma(I)$, as discussed in the introduction.

The third stage involves the transformation of the observations of each reflection into structure factors. This process consists of an averaging step, in which the intensities I of symmetry-related observations and multiply observed reflections are used in a statistical treatment to find the most reliable, averaged reflection values Y. This step implicitly includes absorption and crystal decay corrections. The final result can be expressed in two ways: either as average intensities, called Y, with error estimates $\sigma(Y)$, and their ratio $Y/\sigma(Y)$, or as structure factor F and σ_F. It is the last set of data, consisting of the structure factors F, that is actually used for structure determination. The values for I, Y, and F are all candidates for crystal and diffraction quality assessment.

In the next section each possible descriptor is reviewed. Structural information is derived from F, and immediately related to F is Y. Therefore the Y values will be investigated first.

2.2 Formulation of Criteria for Data Analysis

The formulation of criteria to judge data quality has to be accomplished with practical issues in mind. When structural information is the ultimate goal of data collection, it is essential to obtain a data set as quickly as possible, as completely as possible, and with the inclusion of data at high resolution. This strategy has lead to a data collection technique which implies an arbitrary starting point in data acquisition, as opposed to a technique of indexing a crystal first, aligning it so that chosen reflections are observed, and collecting data from that point forward. From a practical point of view it is undesirable to follow the latter procedure. A second practical point is the amount of time spent collecting a data set. Several days are generally required to acquire all data in a complete data set. Obviously, it is impractical to collect a complete set of data for every crystal whose quality is under consideration, unless the structure is also being determined. These two points, the lack of specific data points chosen ahead of data collection, and the necessary incompleteness of a data set, lead to the following definition of guidelines for crystal quality assessment.

1. Data analysis should be independent of specific data points. In one extreme, it is possible to compare the (100) reflections of each crystal. In the other, it is also possible, and more practical, to compare a set of arbitrary reflections with another set of arbitrary data points, by means of statistical analysis.

2. Data analysis should be independent of how many data points are available in the set. This will allow for comparison of data sets of unequal size.

3. Data analysis should be independent of the redundancy of the data points observed.

If the requirements stated above are too stringent, they can be easily restated in order to be more applicable, but at the same time still useful. The same guidelines are as follows.

1. It is necessary to determine if specific data points are required in a partial data set to make the partial set representative for the entire set and hence for the crystal. If such points are required, they need to be defined before the experiment is conducted.

2. It is necessary to determine if the number of points in the data set influences the apparent quality of the entire set. If this is so, it will be necessary to compare equal size data sets only.

3. It is necessary to determine if the redundancy, whether as an absolute number or as an average redundancy, influences the apparent data quality. If so, it will be necessary to make sure that all data sets contain data points with the same redundancy.

2.3 Reflections and $Y/\sigma(Y)$ Curves

The first test is on averaged intensities Y and $Y/\sigma(Y)$, since the symmetry averaged intensities Y are used to calculate the structure factors F, which ultimately determine the crystal structure. This makes $Y/\sigma(Y)$ curves a logical choice as a descriptor. A significant practical advantage of this choice is that $Y/\sigma(Y)$ values can be manipulated by means of scale factors. The scale factors a_j are used for several purposes: to minimize the average error $< | a_j I - < I > | >$, with $< I > = < a_j I >$, to include a compensation for crystal decay, and to include an implicit absorption correction. When a scale factor is set to zero, the associated intensities I will not be incorporated in the average $< I >$.

2.3.1 Materials and methods. Lysozyme was chosen as model protein in the following experiments. Crystals were grown under standard conditions[2] and also served

as control crystals for the micro-gravity experiments on STS-72 (see section 3.2). Lysozyme was obtained from Seikagaku (Ijamsville, MD, lot# E94Z05). The initial protein concentration was 46 mg/ml as determined by absorbance measurement at $\lambda = 280$ nm, using a value of $A_{281nm}^{1cm} = 26.4$.[45] A 20 mM sodium acetate (Fisher Scientific, Fair Lawn, NJ) solution at pH 4.7 was used as buffer, and a 7% (w/v) NaCl (Fisher Scientific) solution in sodium acetate buffer was used as precipitant. The protein was dialyzed extensively against buffer before use. Crystallization was performed in a standard hanging drop vapor diffusion experiment in a Linbro tissue culture box (ICN Biomedicals Inc., Aurora, OH) as described elsewhere.[36] The drops consisted of equal volumes (15 µl) of protein and precipitant solutions, giving a total initial volume of 30 µl.

The crystal was 0.55x0.4x0.3 mm in size with a calculated volume (assuming cubic shape) of $66*10^{-3}$ mm^3. The space group is[26]: $P4_32_12$ (space group #96) with a unit cell of a = b = 79.58Å and c = 38.13Å. The crystal was mounted in a glass capillary (Charles Supper Co., Natick, MA), which was sealed at both ends with wax. Data collection took place at 20°C on a Siemens multiwire area detector, with copper K_α radiation from a Rigaku RU-200 X-ray generator. The radiation was generated at 40 kV and 100 mA, and passed through a graphite monochromator and a 0.3 mm collimator. The data set was reduced to structure factors with the Xengen package,[28] following exactly the recommendations as given in the manual,[27] with the exception of the q-value for data rejection in program REJECT which was always set equal to seven. All data used for evaluation were taken from the output of the program STATS.

2.3.2 Results and discussion. The objective was to study the criteria as outlined in section 2.2. In order to do so, 360 frames of data, each 0.25° in ω, giving a total of 90°,

were collected. Each frame was exposed for 60 seconds at a crystal to detector distance of 12 cm. A second data set was collected on the same crystal, with identical settings, resulting in a duplicate data set. The individual data sets are 83% complete at 1.77Å, and the average redundancy is 2.2, with $R_{sym} = 6.0\%$. $R_{sym}(I)$ is defined[17] as $\sum_{hkl}\sum_j |I_j(hkl) - \overline{I(hkl)}| / \sum_{hkl}\sum_j I_j(hkl)$. In order to verify the criteria (section 2.2), the data sets will now be evaluated both separately and combined. In addition, subsets of the obtained data set will be compared to the original set.

Figure 5 shows averaged $Y/\sigma(Y)$ values as a function of averaged resolution ($1/d^2$) for data sets A, B, and A plus B combined. The $Y/\sigma(Y)$ curves for set A and B are practically identical. This result is to be expected and it is a necessary condition for a well-behaved quality descriptor. The data in set A and B should theoretically be identical and differ only in experimental error, the decay of the crystal over time, and other changes caused by exposure to X-rays.

When the data sets are combined, the $Y/\sigma(Y)$ curve is significantly different, due to improved statistics. Theoretically this can be explained by assuming that the noise in the data is reduced on average by a factor $\sqrt{N}$, where N is the increase in average redundancy. In the case of the combined data, N is equal to two. Redundancy is defined as the number of observations per reflection and does not need to be a constant over the entire resolution range. However, in this particular example the redundancy must be equal to two as a result of the experimental design. Hence the $Y/\sigma(Y)$ values should increase by $\sqrt{2}$. In order to verify this statement, the $Y/\sigma(Y)$ values from data set A were multiplied by $\sqrt{2}$ and compared with the values from the combined data sets (Table 1). The

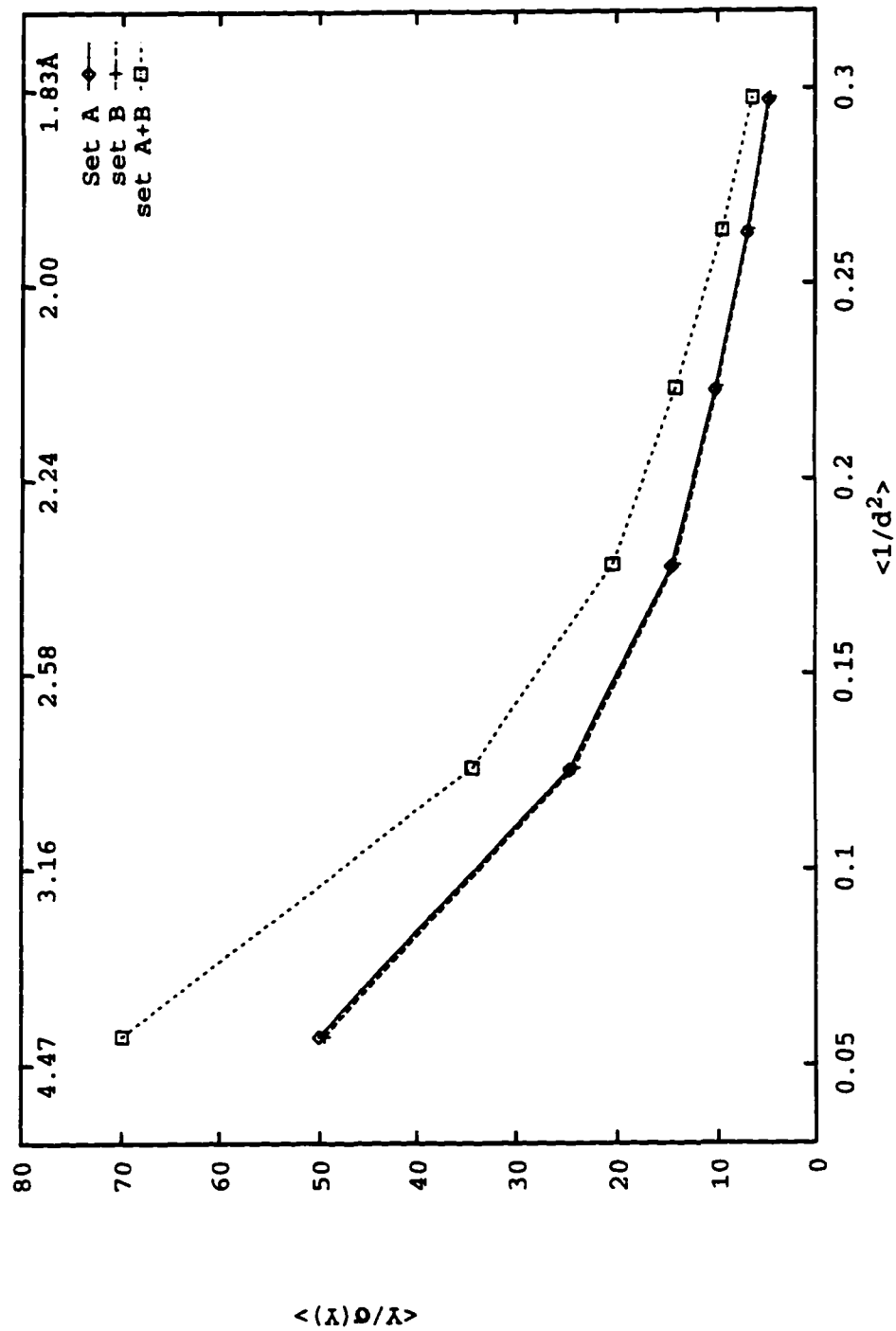

Figure 5: Comparison of $Y/\sigma(Y)$ curves from two identical lysozyme data sets and from their combined data.

Table 1: Comparison of $Y/\sigma(Y)$ from data set A, B, and both A and B combined

average resolution (Å)	$Y/\sigma(Y)$ data set A	$Y/\sigma(Y)$ data set B	$Y/\sigma(Y)$ A+B combined	data set A $*\sqrt{2}$
4.184	50.10	49.51	69.98	70.85
2.823	24.56	24.15	34.34	34.73
2.373	14.63	14.36	20.35	20.69
2.119	10.20	9.98	14.20	14.42
1.946	6.88	6.68	9.49	9.73
1.835	4.66	4.50	6.35	6.59

data presented clearly show that the error between data sets is small and that the combined set can be reliably estimated from either individual set.

In order to judge the influence of the data set size, the data as obtained in set A (Figure 5) are used in the following experiment. The full set, which consists of 90° of data, is cut down by eliminating part of the data. This can be accomplished by setting the scale factors a_j to zero for the data to be excluded. Using this method, the data are represented in bins or shifts, each shift approximately equal to 10 data frames (this is only correct if the frame width is 0.25°).[27] The full data set is equal to 35 shifts. The size of data set A is reduced in steps of 5 shifts (the equivalent of 50 frames per step) and after each deletion new $Y/\sigma(Y)$ curves are calculated. These curves are plotted in Figure 6.

It is obvious from Figure 6 that there is a trend indicating lower $Y/\sigma(Y)$ values when the data set shrinks in size. This result logically follows from the conclusions derived from Figure 4 and Table 1: when the data set size decreases, the redundancy goes down. Since the redundancy is higher for low resolution data (shown in Table 2), the effect is more pronounced on low resolution data than on high resolution data. The data in Table 3 illustrate how the average redundancy decreases with the size of the subset of data.

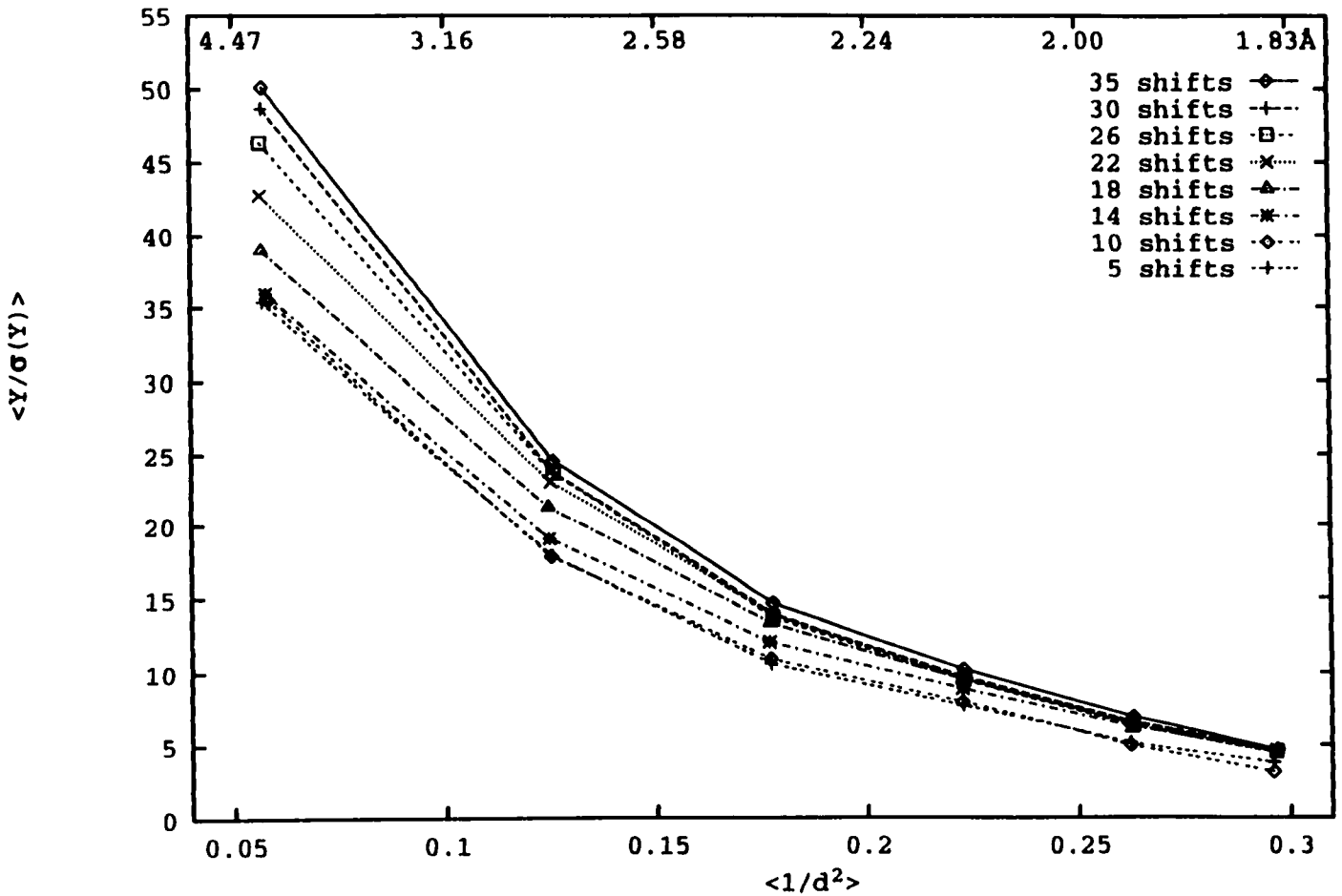

Figure 6: Comparison of *Y/σ(Y)* curves from unequal size data sets from the same crystal.

Table 2: Average redundancy as a function of resolution, for data set A.

shell lower limit Å	average redundancy
3.2131	2.64
2.5502	2.45
2.2278	2.20
2.0241	2.11
1.8790	1.98
1.7682	1.60

Table 3: Average redundancy as a function of number of shifts.

shifts	average redundancy
2-35	2.23
2-31	2.05
2-27	1.92
2-23	1.83
2-19	1.66
2-15	1.40
2-11	1.17
2- 6	1.08

The last step is similar to the previous one, except for the fact that equal size subsets of data are extracted from set A. Each subset consists of approximately 50 frames. The $Y/\sigma(Y)$ curves are determined for each set and plotted in Figure 7. The average redundancy for these subsets ranges from 1.08 to 1.18. During data processing of shifts 7 through 11, it became obvious that the data are unfit for comparison with other subsets. Due to a lack of symmetry related observations, it is not possible to determine scale factors for two out of five shifts. Consequently, these scale factors are zero and the subset contains only three, as opposed to five shifts of data. Figure 7 indicates a more or less constant $Y/\sigma(Y)$ plot, except for shifts 7 through 11 as discussed.

2.3.3 Conclusions for $Y/\sigma(Y)$ curves. From the data presented in Figures 5, 6, and 7 , the following conclusions can be derived for lysozyme:

1. $Y/\sigma(Y)$ curves are remarkably independent of random (non-systematic) errors.

2. $Y/\sigma(Y)$ curves are very sensitive to redundancy. The intensities Y are constant, but the σ-values decrease as the number of observations increases. This means that redundancy is a critically important parameter to keep constant in crystal comparison, or alternatively, that the redundancy has to be explicitly specified with all data sets.

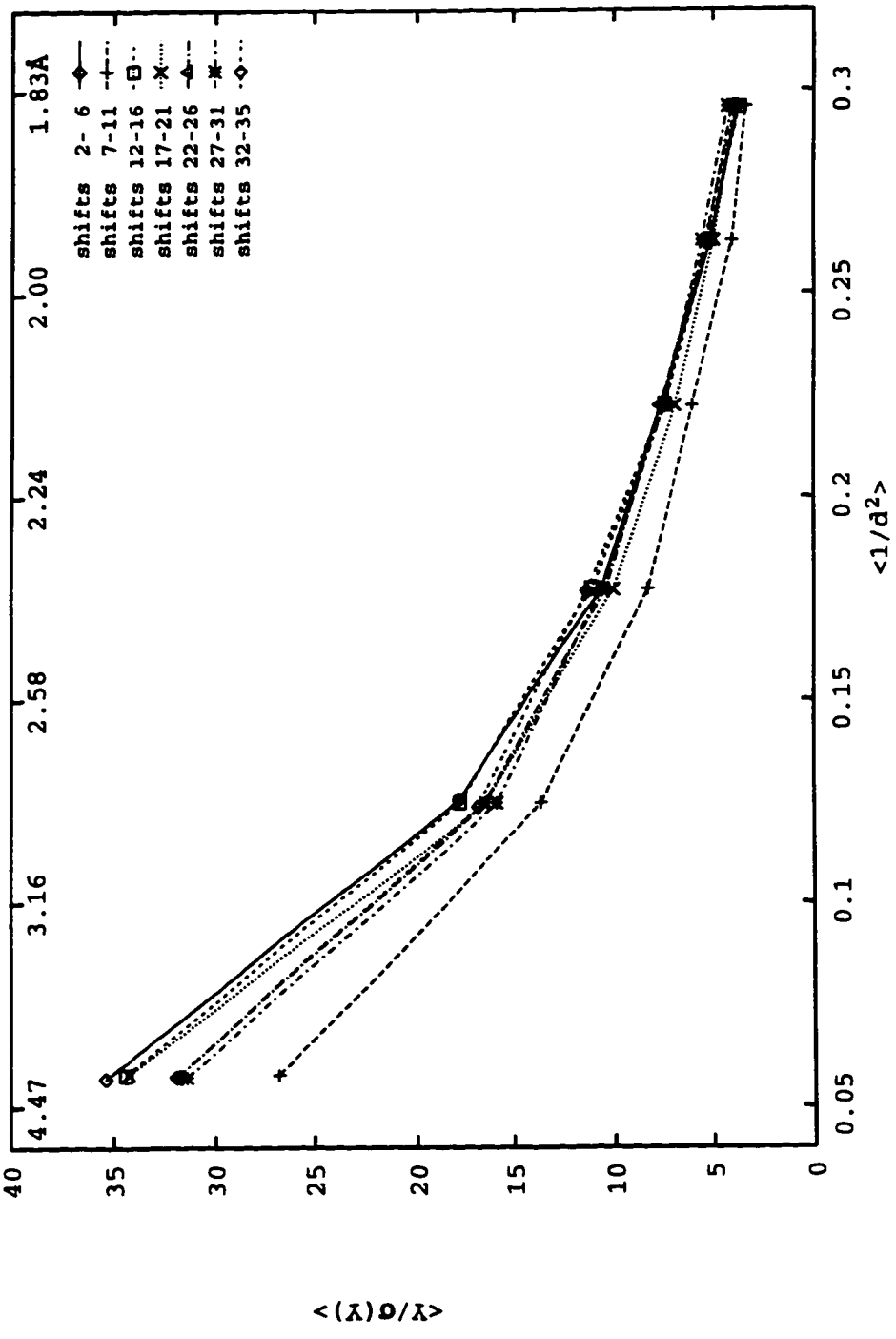

Figure 7: Comparison of $Y/\sigma(Y)$ curves from equal size data subsets.

3. The size of the data set influences the $Y/\sigma(Y)$ curves, probably via the dependence of the average redundancy on data set size. The data set size needs to be constant for useful comparisons of $Y/\sigma(Y)$ values.

4. There is no need for specific data points in a data set, as shown in Figure 7.

2.4 Observations and $I/\sigma(I)$ Curves

The crystal and data used are the same as in the previous section, therefore the materials and methods as described in section 2.3.1 apply to the experiments in this section as well. Instead of using reflections, which are averages of symmetry-related observations, the intensities I are used. The goals are the same: to verify the criteria as set out in section 2.2. Verification will consist of determination of the effect of redundancy, the effect of data set size, and the importance of inclusion of specific data points.

2.4.1 Results and discussion. Figure 8 shows the observations, that is $I/\sigma(I)$, for data sets A, B, and A plus B combined. The $I/\sigma(I)$ curves derived from data sets A, B, and A plus B combined are remarkably similar. Unlike the result for the $Y/\sigma(Y)$ curves, the combined data set does not differ from the individual data sets. This means that redundancy does not influence the statistics of observations.

Figure 9 displays average $I/\sigma(I)$ values calculated for subsets of data set A. The curves in this figure were determined independently of each other. The graph clearly shows that there is no significant difference between data sets of unequal size. There are one or two curves whose shapes are slightly different from the other curves because of a different number of points on the curve, but the points present do coincide with the rest of the curves.

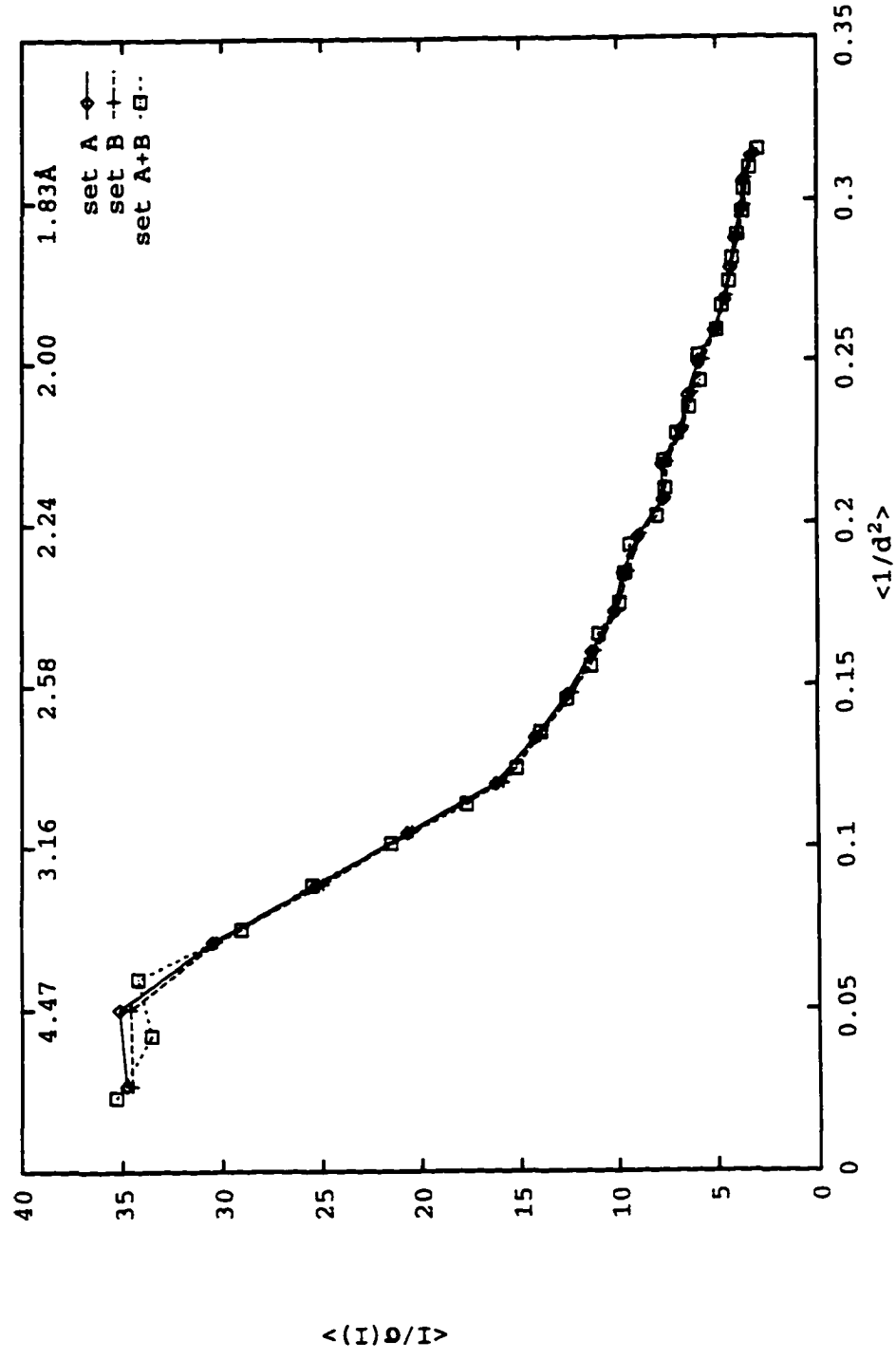

Figure 8: Comparison of $I/\sigma(I)$ from observations in two identical data sets and their combined data.

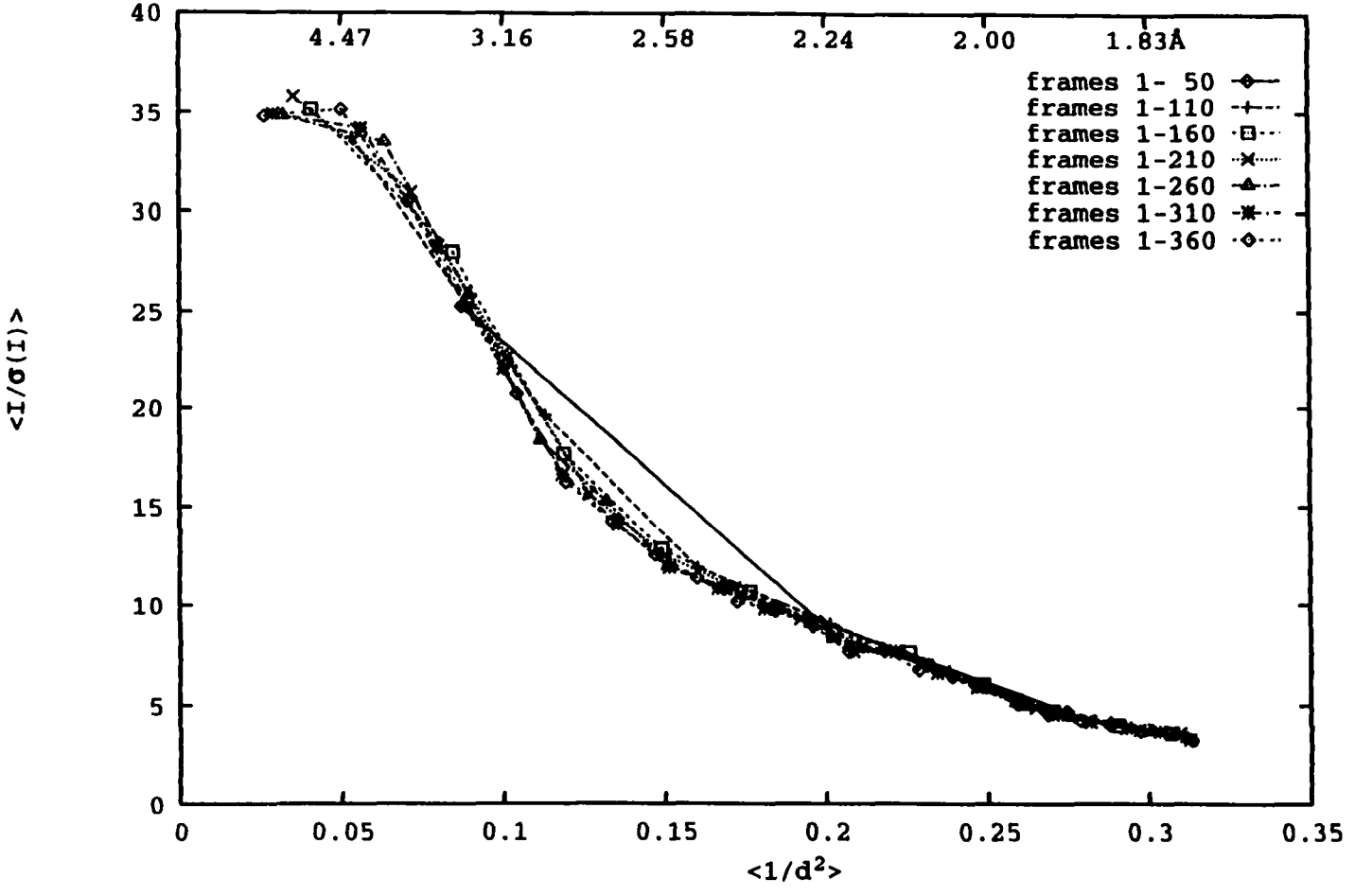

Figure 9: Comparison of $I/\sigma(I)$ from observations in data sets of unequal size.

Finally, in Figure 10, $I/\sigma(I)$ curves derived from 50 frame subsets of data set A are plotted. The curves are very similar, with a slight trend toward smaller $I/\sigma(I)$ values with increasing frame numbers, especially when $1/d^2$ is small. This effect is small but noticeable. Since the data are not corrected for decay or absorption, it is likely that this trend is caused by crystal decay. Other than this small decrease, the curves are very similar to each other and any individual curve gives an adequate description of the entire data set.

2.4.2 Conclusions for $I/\sigma(I)$ curves. Figures 8, 9, and 10 lead to the following conclusions for lysozyme:

1. Redundancy does not change $I/\sigma(I)$ curves.
2. $I/\sigma(I)$ curves do not change with data set size.
3. No specific data points are needed to represent $I/\sigma(I)$ curves.
4. Crystal decay must be considered when using $I/\sigma(I)$ curves.

2.5 B-factors

2.5.1 Method of B-factor determination. As described in section 1.4.2, B-factors can be determined by means of Wilson plots. The Wilson plot consists of $\ln(\overline{I_{hkl}} / \sum_a \overline{f}_a^2)$ as a function of $(\sin\theta/\lambda)^2$. There are several assumptions at the base of this method:[46] the atoms are randomly dispersed in the unit cell; in practice the data are divided into concentric shells in reciprocal space with the understanding that within each shell the resolution $[(\sin\theta/\lambda)^2]$ is constant; and there will be a sufficient number of reflections averaged into each shell to be statistically significant. Starting with Furey's PHASES package,[23] the program CMBISO was modified to produce absolute Wilson plots, as opposed to relative plots.

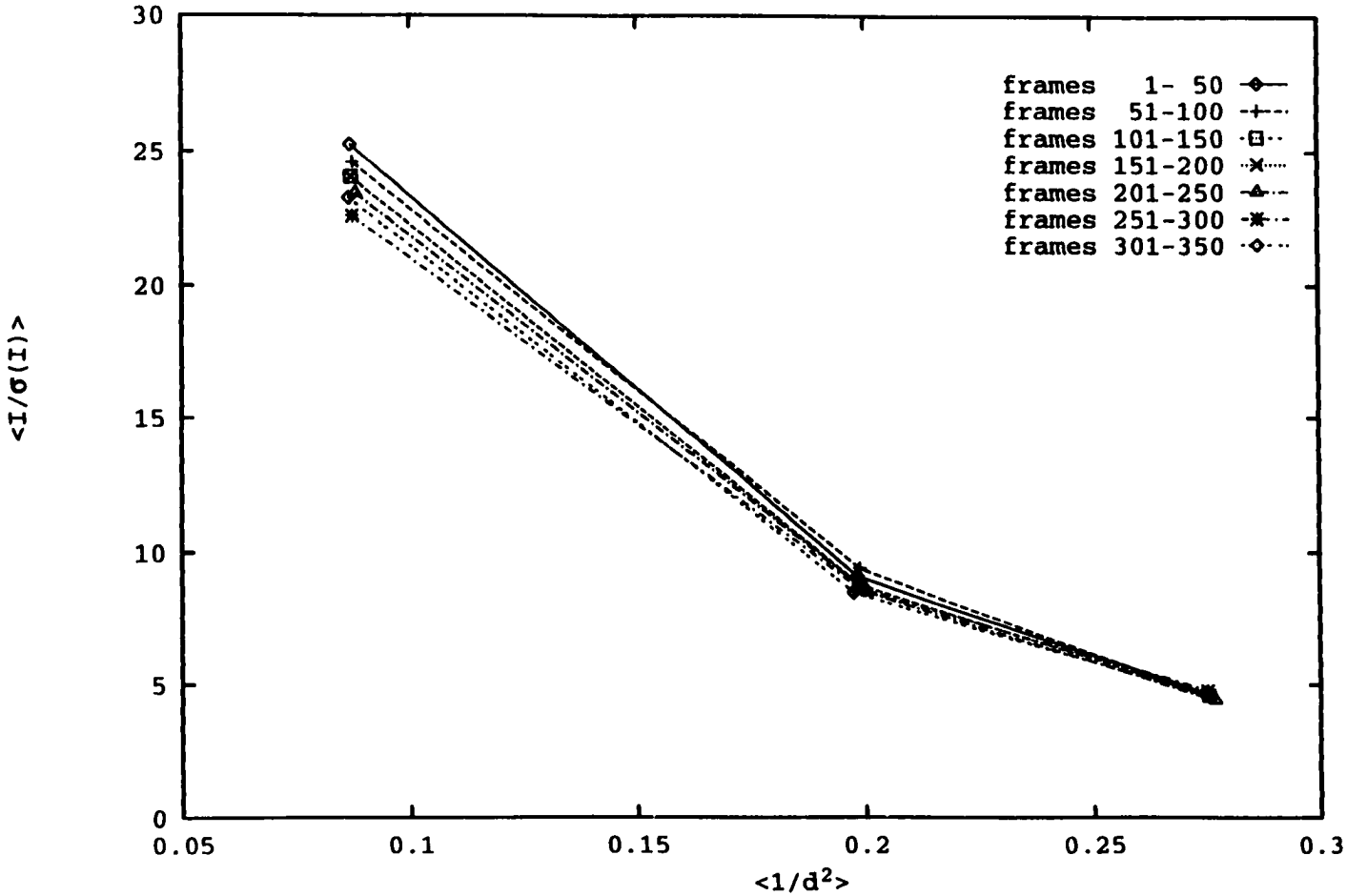

Figure 10: Comparison of *I/σ(I)* from observations in equal size data subsets.

Relative Wilson plots are used for scaling two data sets with respect to each other. The advantage of absolute Wilson plots is that the atomic scattering factors, as discussed in section 1.4.2, serve as a reference point. This reference point is uniquely defined for every structure, as long as the amino acid composition of the protein under investigation is known. From the amino acid composition the exact number of carbon, nitrogen, oxygen, sulfur, and other atoms can be determined and with the available scattering factors, $\sum_a \bar{f}_a^2$ is defined.[29] For lysozyme the amino acid sequence was obtained from the Brookhaven protein data bank, entry 1HEL. Only atoms that are part of the protein molecule were used. The data are typically divided into twenty resolution shells and, for each shell, $\bar{I} / \sum_a \bar{f}_a^2$ is determined. The logarithm of the ratio is plotted as data points in Figure 11 for lysozyme data set A, as used in the previous sections. The data points in Figure 11 lie more or less on a straight line, except for the points at low resolution. The deviation from linearity is caused by the breakdown of one of the assumptions: atoms are not randomly dispersed in the unit cell, since there is a minimum distance between the atoms. This distance is approximately 1.5Å, the bond length between two carbon atoms. Wilson recognized this problem and introduced an estimate for the low resolution cutoff based on statistical analysis.[50] This estimate for s is in the order of a^{-1} to $2a^{-1}$, with a the typical dimension of the unit cell edge. This is the same as a cutoff on $(\sin\theta/\lambda)^2$ of $1/4a^{-2}$ to a^{-2}, which amounts to $6*10^{-4}$Å^{-2} for tetragonal lysozyme. Drenth[17] suggested a cutoff value for s of approximately 1/3Å^{-1}, which translates to a cutoff on $(\sin\theta/\lambda)^2$ of about 0.027Å^{-2}. Observing Figure 11 it appears that a cutoff value of 0.027Å^{-2} is indeed satisfactory. In addition to the data points in Figure 11 there are two lines drawn. One is a

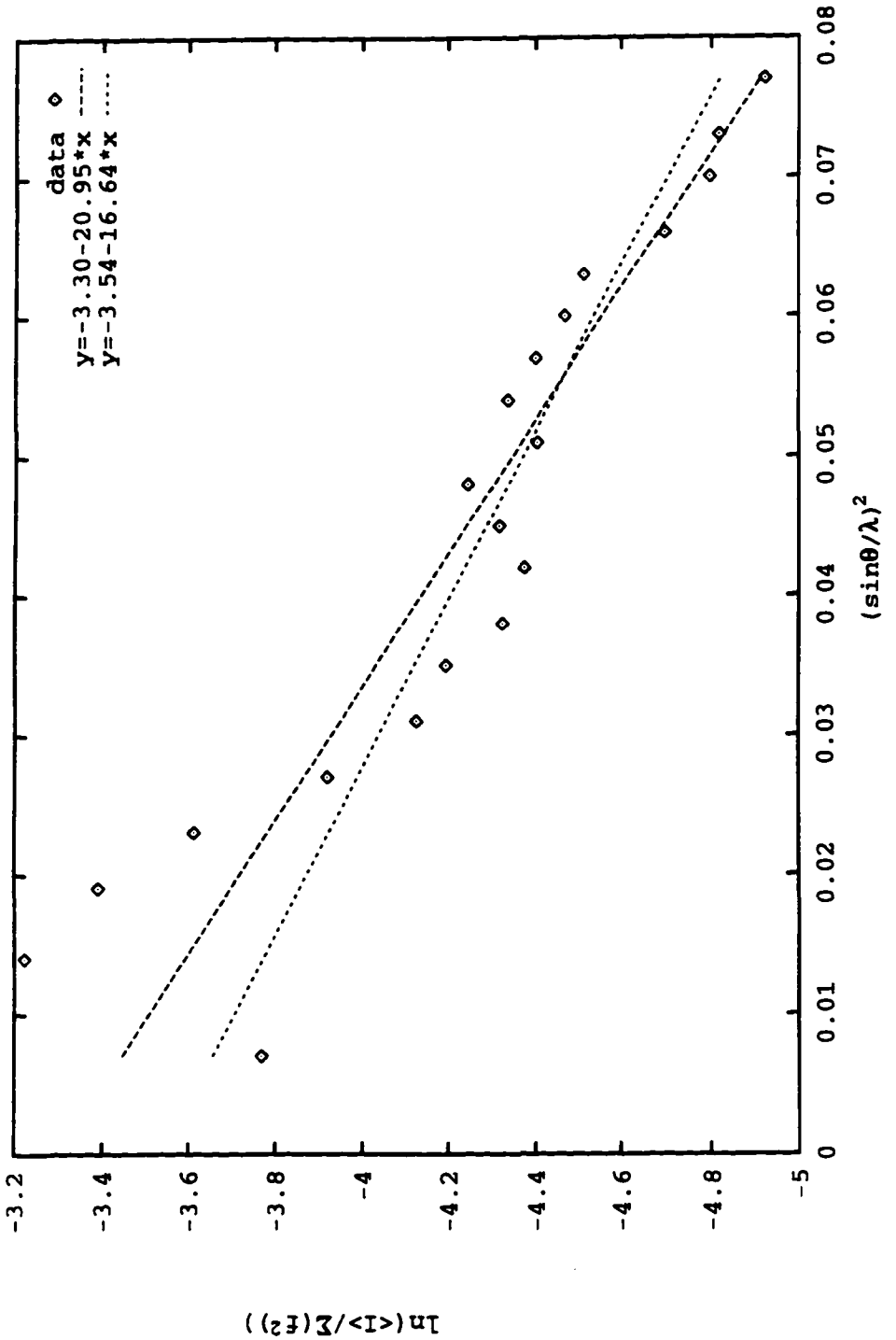

Figure 11: Wilson plot for lysozyme data set A.

least squares fit to all the data points, and the second is a similar fit which excludes all points below the cutoff value of 0.027Å^{-2}. The slope of the line fitted to all data points is steeper than the slope of the second line. The slope of either of these lines is a B-factor estimate. The slope for the line with fewer data points is likely a more reliable value, since it does not include data points for which one of the assumptions does not hold.

2.5.2 Experiments. The main question is whether B-factors are suitable data quality descriptors with regards to the criteria set out in section 2.2. The influence of redundancy, of data set size, and of specific data points on the B-values requires testing, as was done for $Y/\sigma(Y)$ and $I/\sigma(I)$ curves. The lysozyme data sets used in the previous sections are again used in this section. Description of the crystal, its growth condition, and data acquisition are given in section 2.3.1.

2.5.3 Results and discussion. The B-factors are listed in Table 4 for the same lysozyme crystal as used in sections 2.3.1 and 2.3.2. The variable in this table is the number of data frames. The 720-frame data set is actually a combination of the two identical data sets A and B. Therefore the comparison between 720 and 360 frames constitutes a test of redundancy. After the first two entries, the frames are taken from data set A and their number is decreasing. Finally, 50-frame subsets of data are taken from set A and the B-values are calculated in order to determine any dependance on specific data points. Table 4 shows two B-values for each entry, one derived from all data points, and one for data points with a cutoff at 3.16Å resolution. When fitting the data points to a line with the least squares method, a correlation coefficient is obtained for each line. This coefficient has little meaning by itself, but when comparing coefficients, the comparison shows which line is a better description of the experimental data points.

Since the slope of all lines is negative, the correlation coefficients are also negative and a value equal to -1.0 indicates perfect linear behavior.

The B-values in Table 4 are all in the order of magnitude of $10Å^2$. The variation between the entries is remarkably small, which indicates that redundancy, data set size, as well as which subset of data points is used, are all without consequence to the B-value. The values obtained from the data with cutoff at approximately 3Å are somewhat smaller than the associated B-values obtained with all data points. This is consistent with the fact that the curves at lower resolution than 3Å show a maximum. According to Drenth,[17] these high values are caused by the non-covalent interactions in a protein structure, with an abundance of inter-atomic distances of 4Å.

Table 4: B-values as determined by Wilson plots for lysozyme.

#frames	B-value ($Å^2$) all points	B-value ($Å^2$) 3.16 Å cutoff	corr. coeff. all data	corr. coeff. 3.16 Å cutoff
720	11.27	8.98	-0.928	-0.943
360	10.48	8.32	-0.937	-0.947
310	10.51	8.33	-0.927	-0.937
260	10.80	9.06	-0.933	-0.940
210	11.14	8.93	-0.927	-0.952
160	11.12	9.08	-0.923	-0.946
110	7.40	4.78	-0.677	-0.424
1-50	13.37	13.10	-0.900	-0.850
51-100	11.43	8.26	-0.903	-0.831
101-150	10.60	7.76	-0.891	-0.802
151-200	11.28	9.09	-0.925	-0.942
201-250	10.33	7.96	-0.905	-0.880
251-300	9.40	6.49	-0.897	-0.896
301-350	10.22	7.87	-0.915	-0.954

2.5.4 Conclusions for the use of B-factors. B-factors are excellent descriptors of data quality as they are independent of redundancy and data set size. No specific reflections are necessary in the data set to obtain a representative B-factor. Furthermore,

B-values are a concise description of the shape of the $I/\sigma(I)$ curves. A large B-value coincides with a fast decrease of $I/\sigma(I)$ as a function of resolution.

2.6 Analysis of Pixel Values in raw Data Frames

2.6.1 Introduction and background. When solving a crystal structure the first stage in the process of data acquisition is a series of frames with partial or full reflections. Each reflection is represented by a number of pixels in the data frame. Each pixel has an intensity value assigned to it. The steps in data collection and data processing are outlined in Figure 4, and an example of a data frame taken from a lysozyme crystal is given in Figure 12. Carson developed a new method to assess data quality from frames like these.[11,33] This method does not include any of the steps outlined in Figure 4. A short description of the method is as follows: a number of frames of data, usually 15, are merged into one frame by taking the maximum value observed at each pixel. Based on equal increments of $1/d^2$, the summed data frame is then divided into ten bins of resolution. The background level in each bin is determined separately in two passes. The root mean square values and standard deviations of the pixel values are calculated, and those pixels with counts greater than 1σ above the root mean square average are discarded. The remaining pixels are used to recalculate the root mean square average and standard deviation for each bin. The average intensity $\bar{I}$ of probable reflection pixels is evaluated by summing the values of pixels greater than 3σ above the root mean square average and dividing by the number of pixels in the bin. The result of this analysis is a graph of $\bar{I}/\sigma$ as a function of $1/d^2$, with $\bar{I}$ and σ as just defined.

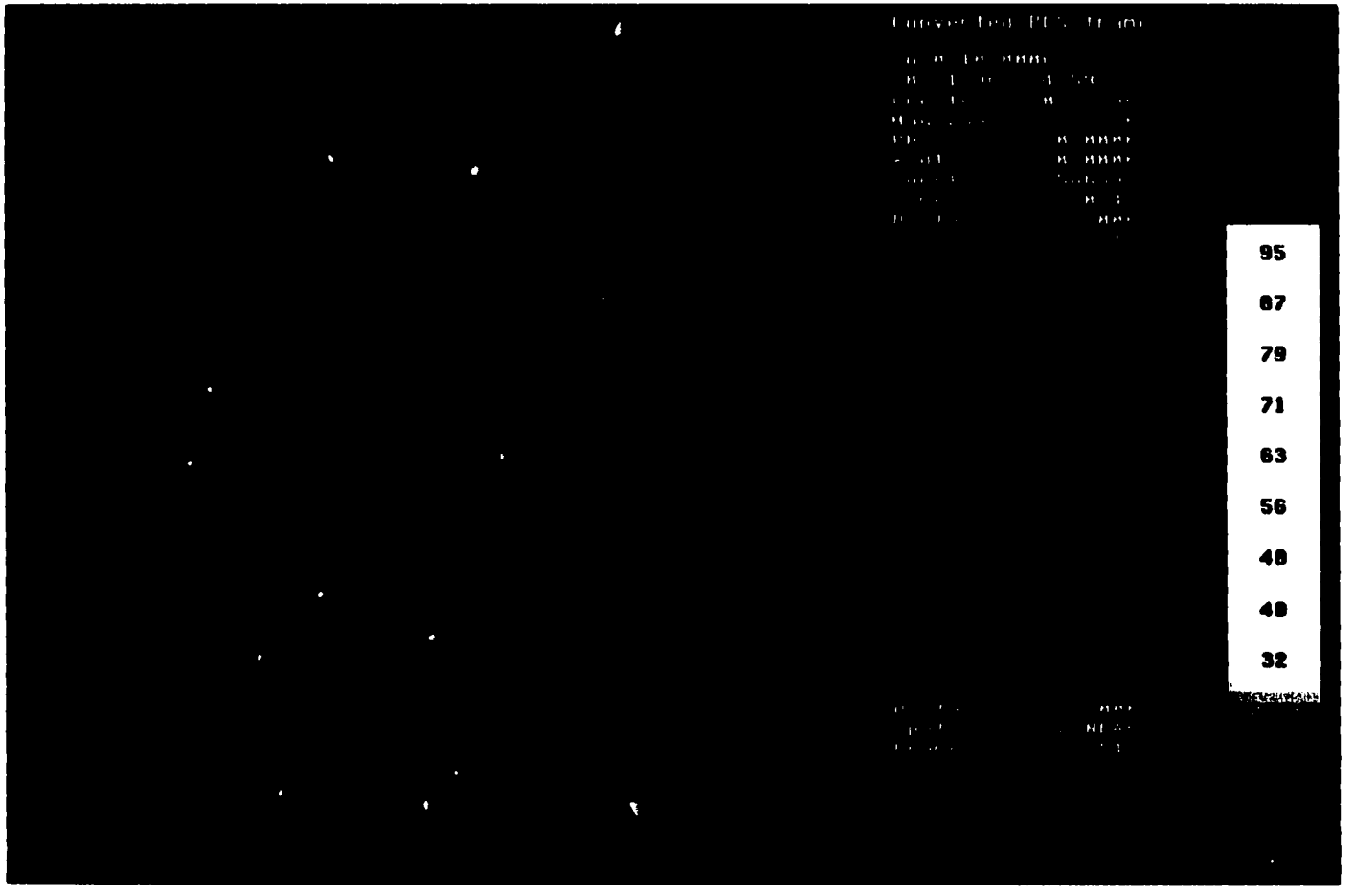

Figure 12: An example of a raw data frame as acquired in the first step of data collection.

2.6.2 Objective. Following the method presented in the previous section, it would seem obvious to test the data based on pixel values versus the criteria set out in section 2.2. Since this is a new method, however, the first step is a comparison of data interpreted with this new method versus standard intensity and background determinations. This ensures that the data representation with the new method contains the same information as the standard representation.

2.6.3 Data interpretation. Comparison of standard $I/\sigma(I)$ curves with the new method based on pixel values was accomplished based on data made available by Dr. Craig Smith. The objective of the experiment leading to this data was to compare the quality of data from crystals examined at room temperature, versus data from frozen crystals. The protein crystals used were ribonuclease-S (bovine pancreas, grade XII-S, Sigma Chemical Company, St. Louis, MO) crystals provided by Vickie King. Crystals were grown following a literature procedure,[51] with a ribonuclease-S solution of 75 mg/ml (weighed) in 6 M CsCl (J.T. Baker, Phillipsburg, NJ) and a precipitant of 95% (v/v) saturated ammonium sulfate solution and 5% sodium acetate solution. The saturated ammonium sulfate solution was prepared by dissolving 153.4 g $(NH_4)_2SO_4$ (Fisher Scientific) in 200 ml of deionized water. The sodium acetate solution is a 0.1 M $NaOCOCH_3 \cdot 3H_2O$ (Fisher Scientific) solution, with a pH adjusted to 6.1 by adding HCl. The crystals belong to space group $P3_121$, space group[26]#152 , with cell dimensions of approximately a = b = 44.7 and c = 97.0Å. The crystal to detector distance was 12 cm. Data were processed with the Xengen package.[28] The standard $I/\sigma(I)$ curves are plotted in Figure 13. There are two main features visible in this graph: first, the curves labeled RNS1123A through RNS1123E are similar to each other, but distinctly different from the

curves labeled RNS1201 and RNS1201B; second, the curves labeled RNS1201 and RNS1201B are similar in that they contain data points in the same resolution range. Several observations can be made: data collected at room temperature have lower signal-to-noise ratios than data collected from a frozen crystal, and the data quality is independent of the detector settings, for as far as data points overlap. The change in detector settings in this experiment is a variation in 2θ value, which changes the resolution limit in the data set.

Figure 14 displays the information as obtained with the new method outlined in section 2.6.1, from the same crystals. The data analyzed in this manner are very different from Figure 13. The conclusions reached from data in Figure 14 would be that there is no difference between data from the various crystals, and that detector settings, which are influencing the maximum resolution observable, do indeed influence the average I/σ values for data points at low resolution. These conclusions are incompatible with the conclusions reached from the data in Figure 13. It is obvious from the comparison of these two figures that the pixel analysis does not resemble analysis based on true integrated diffraction intensities. Therefore this new method is neither a replacement for, nor an improvement on the standard diffraction analysis methods.

2.7 Conclusions and Recommendations for Data Quality Assessment

From all the tests described in this chapter, it appears that there are two reliable data quality descriptors: $I/\sigma(I)$ curves and B-factors. It is essential to make sure that the intensities I are observations and not averaged reflections. When data processing packages other than Xengen are used, this difference may not be obvious. In case the intensities are symmetry averaged observations, it is critically important to report the

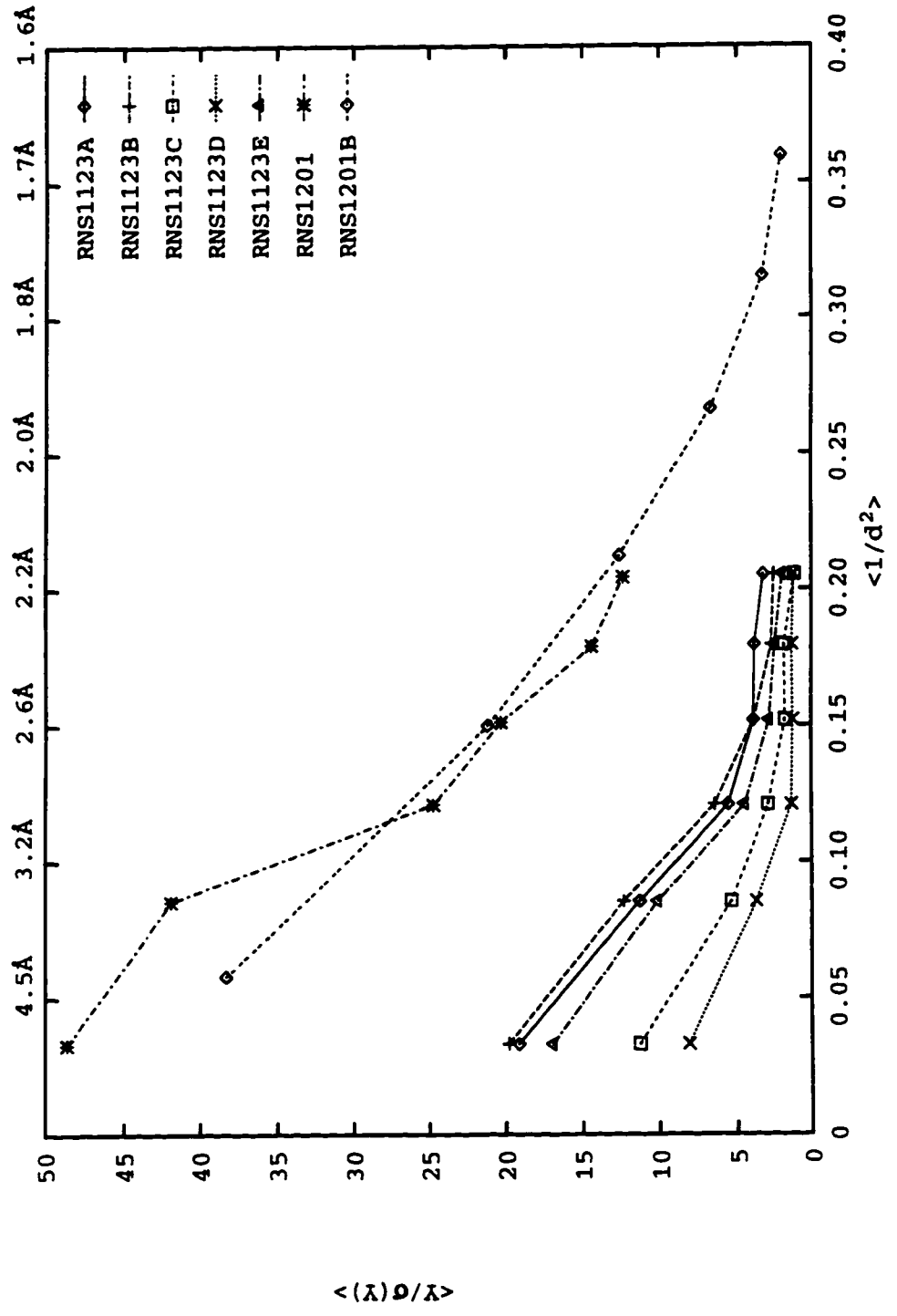

Figure 13: $Y/\sigma(Y)$ plot for various Ribonuclease-S crystals.

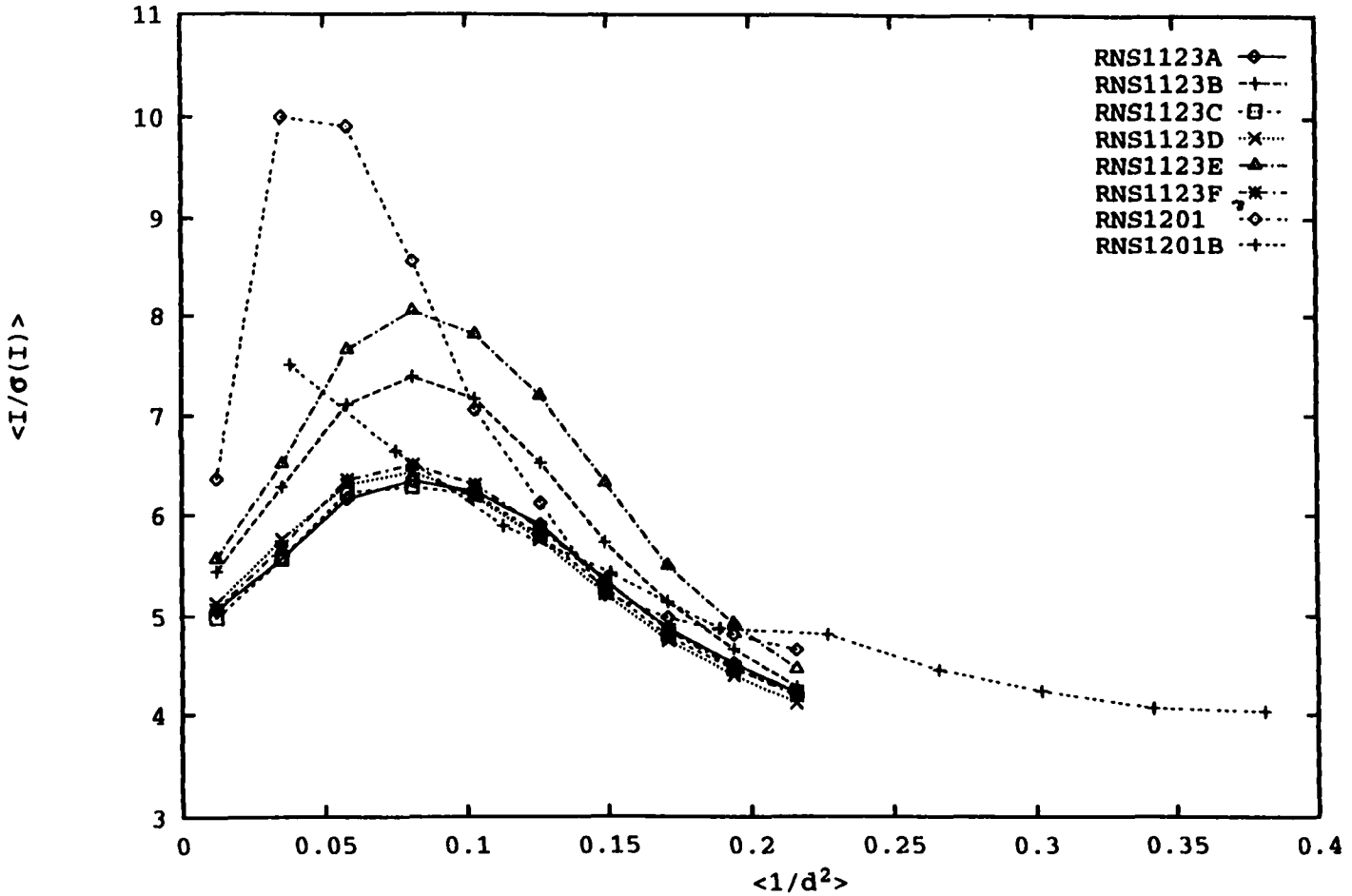

Figure 14: $\bar{I}/\sigma(I)$ values as function of resolution, as determined by pixel analysis, for Ribonuclease-S.

multiplicity of the data, if possible, for each resolution bin. These multiplicities must be approximately equal for the compared data sets. B-factors seem to be a more concise descriptor and give an idea about the shape of $I/\sigma(I)$ curves. They probably relate to refined B-factors, but the exact relationship is not obvious at this point. The combination of $I/\sigma(I)$ curves and B-factors describes the maximum resolution at which useful data are available and gives an indication of internal order in the crystal. With the aid of a lysozyme data set, it has been shown that there is no need to collect a full data set, nor are there requirements for data set size or contents: an arbitrary "slice" of data is adequate. Although the experimental set-up was chosen without regard to space group, this is not necessarily proof that this conclusion applies to every crystal, regardless of which space group it belongs to. It is, however, likely that this conclusion does indeed extend to other crystals.

2.8 Verification of Conclusions

The conclusions reached in the previous sections are based on data interpretation from one hen egg white lysozyme crystal. Except for the necessary symmetry considerations during data processing, the analysis has been performed without assumptions based on space group. It is possible that part or all of the results are influenced by coincidental circumstances, notably circumstances governed by space group symmetry. Tetragonal lysozyme crystallizes[2] in space group $P4_32_12$. When experiments similar to those described in previous sections are performed with a crystal in a different space group, the combined conclusions will more likely be relevant for a large population of crystals. An adequate solution is to consider crystals of recombinant human insulin, that crystallize in space group R3. Such crystals can be grown with the formulation as published by Smith

et al.[44] Crystals in space group R3 are completely different from those in space group $P4_32_12$. Differences include crystal class (3 versus 422) and crystal system (trigonal versus tetragonal).[5] In the following sections a description of experiments and data interpretation will be given, aiming at a repetition of procedures used for lysozyme. The goal is to verify if $Y/\sigma(Y)$ curves, $I/\sigma(I)$ curves, or B-factors are adequate quality descriptors.

2.8.1 Materials and methods. The insulin crystal used was made available by Vickie King and was grown using the recipe published by Smith et al. for the $T_3R_3^f$ formulation.[44] Human recombinant insulin was generously made available by Lily Research Laboratories (lot#528KKO) and standard chemicals, hydrochloric acid (Fisher Scientific), zinc acetate (Sigma Chemical Company), sodium citrate (Fisher Scientific), 4-hydroxy-benzamide (Aldrich Chemicals, Milwaukee, WI), and sodium chloride (Fisher Scientific) were used as purchased. A 7 ml sample of protein solution was made following the published recipe and heated to 55°C until a clear solution was obtained.[44] The solution was then cooled to 45°C over one hour, at which temperature the solution was filtered through a sterile 0.22 μm pore size filter (Micron Separations, Westborough, MA). The solution was then allowed to cool to 40°C, after which it was gradually cooled to 22°C over 24 hours. The crystal was chosen based on optical quality and size (0.2 x 0.2 x 0.15 mm). Data collection and processing procedures were completely analogous to the procedure in section 2.3.1, with the exception of the exposure time, which was 120 seconds per frame.

2.8.2 Reflections and $Y/\sigma(Y)$ curves for insulin. The crystal was indexed in space group R3, hexagonal setting, space group #146,[29] with unit cell parameters a = b = 80.7

and c = 37.7Å. Two data sets of 360 frames each were collected with a frame width of 0.25° (ω scan). The second set was collected with exactly the same parameters as the first set, in order to study the influence of redundancy on the data. The first data set is 68% complete to 2.12Å, with an average redundancy of 1.27 and R_{sym} = 3.51%. The second data set is 69% complete to 2.13Å with an average redundancy of 1.30 and R_{sym} = 3.30%. The combined data sets form a data set which is 67% complete to 2.12Å, with an average redundancy of 2.51 and R_{sym} = 4.31%. In Figure 15 the average $Y/\sigma(Y)$ values are plotted as a function of $1/d^2$, for both individual data sets A and B, and for the combined data set, labeled A + B. The curves for data sets A and B are similar, but not exactly the same. Particularly at high resolution they differ. This difference is likely caused by the fact that the data sets are not exactly the same, in spite of efforts to repeat the experiment as carefully as possible. During recording of data set A, the crystal slipped. From optimization of the orientation matrix during integration, once every 50 frames, the movement can be found to be $\Delta\omega = +0.99°(+0.18°)$, $\Delta\chi = +0.25°$ (- 0.01°), and $\Delta\varphi = -0.16°(0.02°)$. The numbers in parentheses are the corresponding values for data set B, in which case no significant crystal movement occurred. It is likely that the $Y/\sigma(Y)$ values in data set A are lower due to crystal movement. The combined data show higher $Y/\sigma(Y)$ values, caused by increased redundancy.

The second test aims at assessing the influence of data set size on the $Y/\sigma(Y)$ values. Data set B, which is better defined than A, was taken and its size was gradually decreased. This was accomplished by reintegration of the frames, while systematically decreasing the number of frames in the data set by about 50 frames per step. The first four samples, 360, 310, 260, and 210 frames did not present problems. In the case of 160

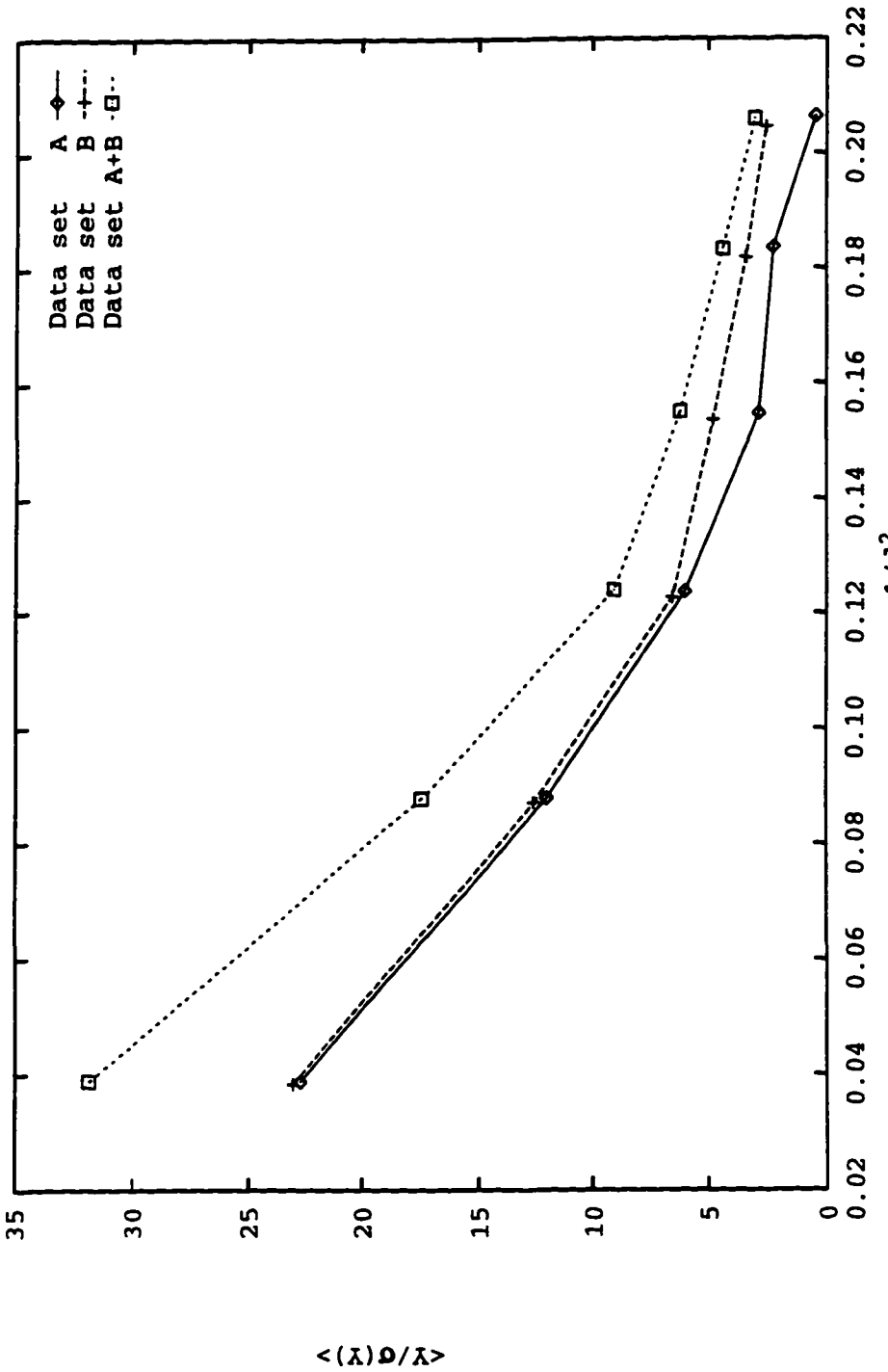

Figure 15: $Y/\sigma(Y)$ plots for human recombinant insulin ($T_3R_3^f$).

frames (14 shifts), 2 out of 14 scale factors are zero. The scaling program reports zero observations for comparison in such a case, which means that no equivalent observations can be found in the frames used and the scale factors are set to zero. In the case of 110 frames, resulting in 10 shifts, there are seven shifts with zero scale factors for the same reason. Finally, in the case of 50 frames, there are no observations available for comparison and all scale factors are zero. In the plot of $Y/\sigma(Y)$ as a function of $1/d^2$, shown in Figure 16, the different curves for various data set sizes are drawn. It is clear from this graph that $Y/\sigma(Y)$ curves do not depend on data set size, until the size decreases past the point where scale factors cannot be determined. This outcome suggests several conclusions: first, since the minimum data set size for lysozyme is equal to or less than 50 frames, this minimum size is in fact space group dependent; second, inadequate data set size, which means too few frames in a data set, can be recognized in normal processing procedures following the Xengen manual by inspection of scale factors.[27] If scale factors with zero value occur, comparison of data sets with apparently equal numbers of frames fails.

The results of a third test, to determine if specific data points are necessary in a subset of data, would be insignificant for the data available for insulin. The total amount of data available is too small to allow for even two non-overlapping subsets of approximately 20 shifts each. This test is therefore omitted.

2.8.3 Observations and $I/\sigma(I)$ curves for insulin. Following the previous section it is important to remember that $I/\sigma(I)$ curves give a different description of the data than do $Y/\sigma(Y)$ curves. Results in this section should therefore be compared to the similar results for lysozyme (sections 2.4.1 and 2.4.2). In Figure 17 the $I/\sigma(I)$ values are plotted for

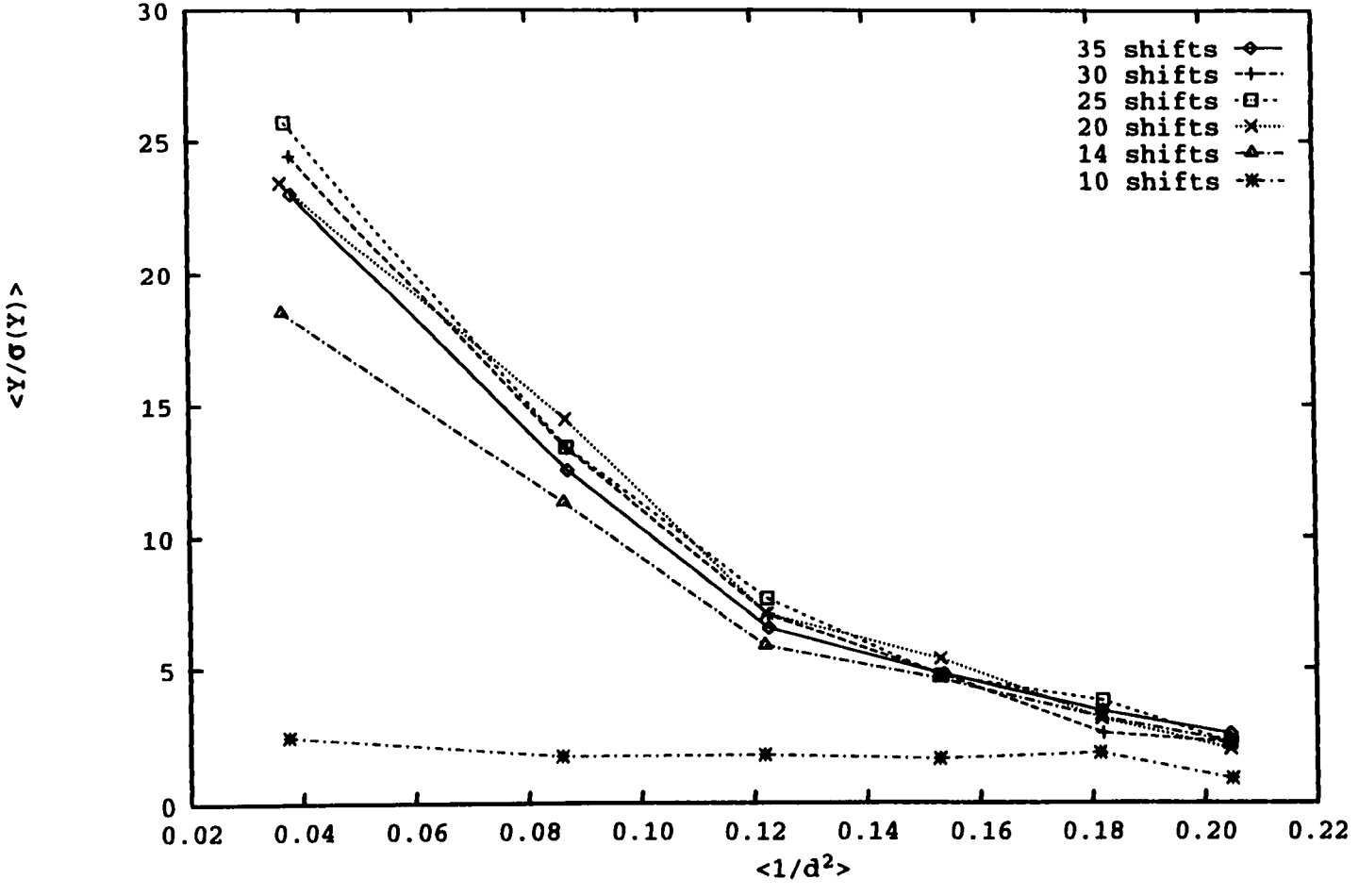

Figure 16: $Y/\sigma(Y)$ plots for human recombinant insulin ($T_3R_3^f$), with variable data set size.

separate data sets A, B, and A plus B combined. Figure 17 shows that the $I/\sigma(I)$ curves are not sensitive to redundancy, since the combined data gives essentially the same curve as the individual data sets. This result is in complete agreement with the results for lysozyme. In Figure 18, $I/\sigma(I)$ curves are plotted for subsets of data set B, with decreasing subset size. The curve with label 35 shifts represents the full data set. Since one shift corresponds to 10 frames, the curves drawn represent data sets decreasing in steps of 50 frames. The curves in Figure 18 are all similar, with the possible exception of the line labeled with 10 shifts, which appears to be somewhat different from the rest of the data. This result may be explained by assuming that the data subset for 10 shifts has become too small to properly represent the full set, as discussed in section 2.8.2. Subsets larger than 10 shifts provide a roughly identical result.

In light of the conclusion at the end of section 2.8.2, it is not reasonable to divide the existing data set into small subsets to test for the importance of specific data points.

2.8.4 B-factor determination for insulin. The method to calculate B-factors for insulin is analogous to the method described in section 2.5.1 for lysozyme. In the case of insulin, the reference point for $\sum_a \bar{f}_a^2$ is the Brookhaven protein data bank entry 1TYM. The amino acid sequence from this entry can be used to determine the approximate atomic composition of the protein with the aid of the Awk program in Appendix B. Following the entry 1TYM, insulin chains A through D are present in the asymmetric unit; the chains form two insulin monomers, each consisting of two heterochains. Water, chloride, zinc, and 4'-hydroxyacetanilide (in 1TYM) or 4-hydroxybenzamide (formulation by Smith et al.,[44] followed to prepare crystals) are not accounted for in calculating

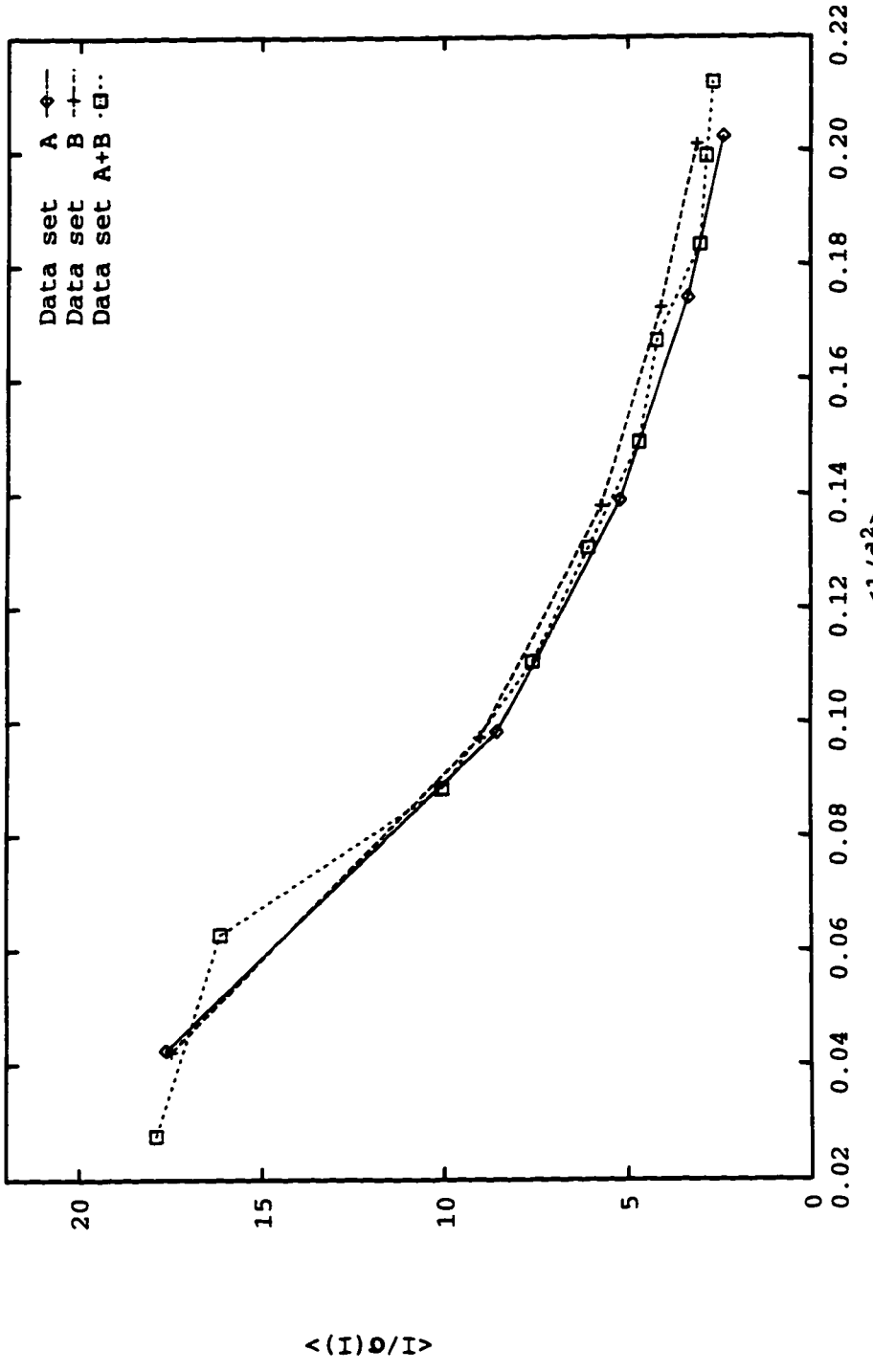

Figure 17: $I/\sigma(I)$ plots for human recombinant insulin ($T_3R_3^f$).

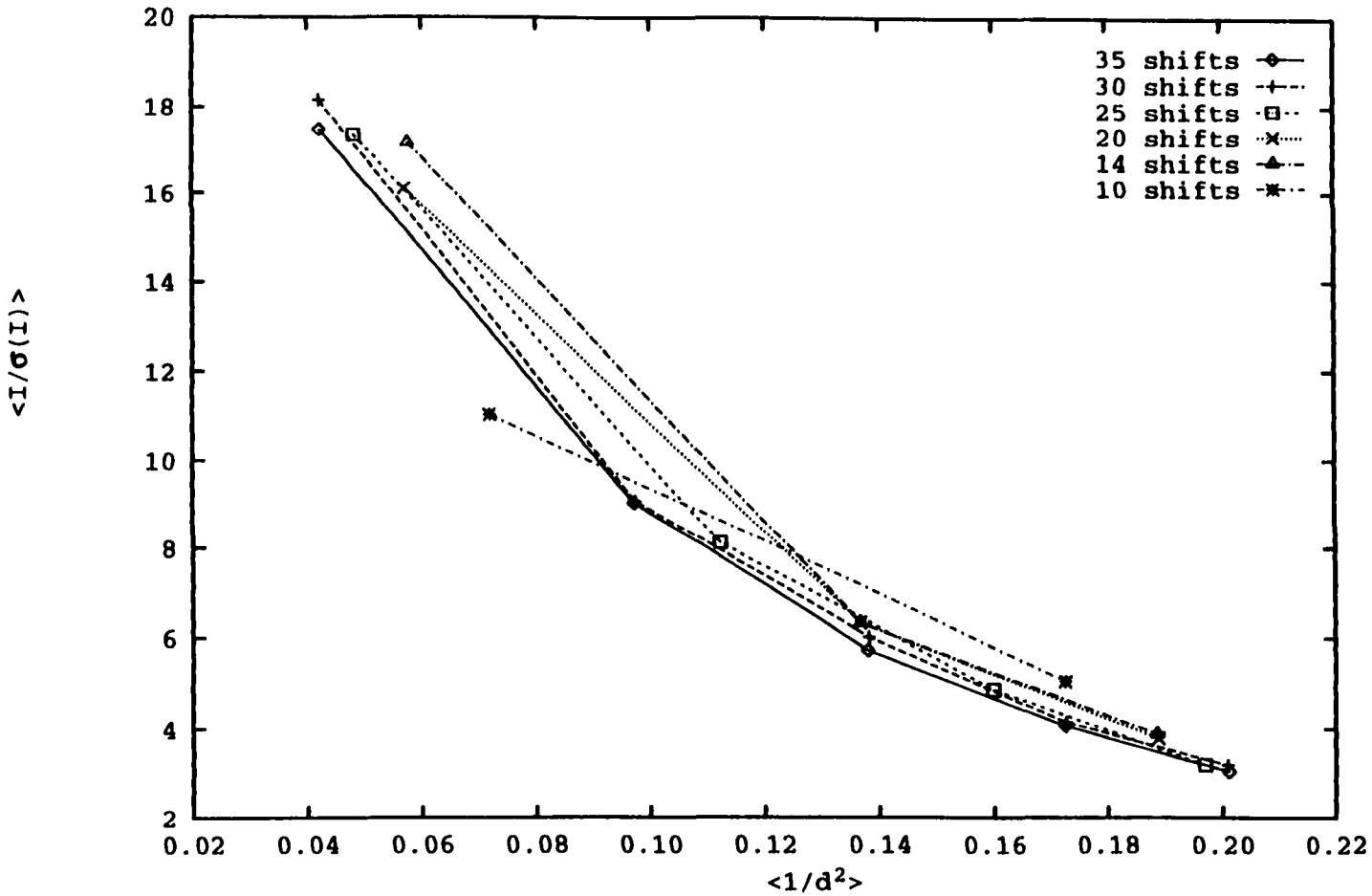

Figure 18: *I*/σ(*I*) plots for human recombinant insulin ($T_3R_3^f$), with decreasing data set size.

the atomic contents of the unit cell. The B-factors for insulin are compiled in Table 5. For each entry, two B-factors are given: one derived from all available data points, together with the appropriate correlation coefficient, and one from data points with $(\sin\theta/\lambda)^2$ greater than 0.027Å^{-2}, together with the applicable correlation coefficient.

Table 5: B-values as determined by Wilson plots for human recombinant insulin, $T_3R_3^f$ formulation.

# frames	B-value (Å^2) all points	B-value (Å^2) 3.16 Å cutoff	corr. coeff. all data	corr. coeff. 3.16 Å cutoff
720	25.08	22.45	-0.974	-0.969
360	21.94	14.68	-0.924	-0.787
310	25.72	24.48	-0.952	-0.927
260	21.49	14.94	-0.874	-0.619
210	24.22	20.41	-0.921	-0.813
160	26.66	23.81	-0.920	-0.887
110	12.08	11.65	-0.633	-0.419

2.8.5 Results and discussion for B-factor determination for insulin. As for lysozyme, the 720-frame data set is actually a duplicate 360-frame data set and can be used to judge the influence of redundancy. The entries in Table 5 run from 720 frames down to 110 frames. Since it is not possible to obtain reliable *Y* values for data sets smaller than 110 frames (see section 2.8.2), it is also not possible to calculate structure factors and hence B-values. The B-factor values are generally constant, with the exception of the case of 110 frames of data. These results support the previous conclusions for the use of B-values in section 2.5.4. This specifically means that B-values neither depend on redundancy in the data set nor on the data set size, given that the set is larger than a certain minimum value. This minimum is apparently not a constant independent of protein and space group. As is the case for lysozyme, the correlation coefficients seem to be a reasonable control for B-factor reliability: low (absolute) values indicate poor

results. In a comparison with the $I/\sigma(I)$ curves shown in Figure 18, the correlation coefficient in the B-factor calculations drops significantly at the same point as the $I/\sigma(I)$ curves start to change.

2.9 Improved Pixel Analysis

2.9.1 Motivation. The conclusion obtained by comparison of true, integrated reflection intensities and pixel analysis as performed in section 2.6.3 indicates that integrated values are essentially different from pixel values. Since integrated intensities lead to structure factors and ultimately to the resolution of protein structures, true intensities must logically represent the data. In the process of obtaining integrated intensities, there are many possible causes for failure. In those cases no final, integrated intensities are obtained. The main problem usually is the inability to index the data. This is to say, no (hkl) values can be assigned to individual reflections. For example, this situation will occur for crystals that are not single. A lack of information about space group or unit cell parameters may also cause problems, as space group and unit cell parameters are a necessary requirement for data indexing. Whatever the reason is for a failure to index data, no information other than subjective information will be available for the particular crystal and data set. An example of such information is the diffraction limit observed by the operator. This analysis is obviously least desirable because it is not objective. This dilemma causes a need for a reliable means of analysis, even if the standard procedures do not work. Data processing speed is another parameter that can lead to an improved analysis method, as pointed out in the literature.[11,33] Initial data processing is never the rate limiting factor in protein structure analysis, but does require significant human interaction. This means that computers cannot perform this task independently. If many

crystals need to be analyzed in order to judge the influence of crystallization conditions on crystal quality, the entire process of obtaining integrated reflection intensity values can become prohibitively time consuming. Both reasons, time and the fact that not all crystals qualify for the standard procedures, call for an alternate approach.

2.9.2 New method. The program SPOTS is used in the first step in data processing with the Xengen package.[28] This program is written in C and consists of approximately 30 routines. The source code spans approximately 30 single spaced pages of text. Considering the size and complexity of this program, only those facts pertinent to the pixel analysis as appropriate to this work, will be discussed. A short description of the program taken from the Xengen manual,[27] SPOTS entry, is:

> *Spots* reads a series of contiguous data frames and constructs from them a list of bright spots to use in refining the crystal and detector parameters. It computes a centroid in detector coordinates (X,Y) and in spindle angle (φ) for each of these reflections and writes the resulting list to disk.

Input for SPOTS consists of data frames, their starting and ending sequence number, and a cutoff value for initial observation of pixel intensity in terms of background value. By default, SPOTS will process all available data frames and use a cutoff value of 15σ. These default values are likely applicable in the majority of the data sets encountered. SPOTS scans squares of 64 x 64 pixels in a data frame, searching for pixels that exceed the background cutoff value. It then proceeds, when such a pixel is found, to search the maximum pixel value in the reflection to which the original pixel belongs, determines in all three dimensions X, Y, and φ which pixels belong to the reflection, sums their values, and subtracts a background which consists of an average pixel value in the same 64 x 64 pixel square in adjacent frames, where no reflections are present. There

are several additional requirements for a well-behaved reflection, such as a minimum integrated intensity, and the reflection cannot be two-humped (any two-dimensional cross-section of the reflection cannot show more than one maximum). When the reflection qualifies, the program writes its X, Y, φ, and intensity values to a binary file, called the centroids file. This file, in turn, is used by REFINE, but more important for pixel analysis, can be read and printed with the aid of the program XDUMP.

The program described above almost provides a reasonable candidate for data analysis. The only item missing is a σ-value in the centroids file. Since SPOTS calculates a background value and uses it in intensity determination, the program SPOTS was modified to include σ-values in the output (centroids) file as follows:

1. change in XDEFS.H "#def SIZREF 20" to "#def SIZREF 24" to reflect the addition of the sigma value in the output records;

2. add ptroi.sigma to the ptroi structure in XDEFS.H, the value of ptroi.sigma will be equal to peakp→psigma and the entire ptroi structure is written to the centroids file;

3. in file SPOTPROC.C, procedure troidcalc, add right after calculation of peakp→psigma "peakp→ptroi.sigma = peakp→psigma;";

4. change in file SPOTPROC.C, procedure troidcalc the value of SIZOUTREF from 20 to 24;

5. add in file XDUMP.C, in procedure prtroid(fp), tr.sigma and tr.totintent/tr.sigma, in order to read the sigma value back from the centroids file and to calculate the I/σ value;

6. in file XDUMP.C, procedure dodump(type,name), for case 3, change headers to include σ and I/σ.

Both SPOTS and XDUMP need to be recompiled after these changes have been made. Since essential structures have been changed in XDEFS.H, the entire suite of programs should be recompiled in order to keep them consistent. After running the new SPOTS and XDUMP programs, a list of reflections is available. A new program made with the utility Awk, a standard utility in most UNIX computer operating systems, reads the XDUMP results and extracts necessary information for $I/\sigma(I)$ graphs. The program is listed in Appendix A. It needs, except for the XDUMP results, the position of the direct X-ray beam at zero detector angle ($2\theta = 0$), the crystal to detector distance in millimeters, the wavelength of X-rays (usually 1.5418Å, for CuK_α radiation), and the detector swing angle 2θ. The program divides the detector surface in ten bins of equal size, based on $1/d^2$ and assigns each reflection to a bin, keeping track of average I/σ values. This average is reported after completion with the average $1/d^2$ value for the bin, the range of $1/d^2$ for the particular bin, and the number of reflections in the bin.

2.9.3 Results and discussion for ribonuclease-S. All materials, methods and procedures as discussed in section 2.6.3 apply, except that the final data points are obtained with the programs described in the previous section. The two points of reference are Figures 13 and 14. The former has been determined by complete data processing with the Xengen package, the latter with the pixel analysis as described in the literature.[11,33] Data analysis for data sets labeled RNS112A through RNS1123E, RNS1201, and RNS1201B is shown in Figure 19. For each data set, 60 frames were processed with a cutoff value of 15σ. Figure 19 indicates no significant difference in data in the series RNS1123A through RNS1123E, whereas RNS1201 and RNS1201B appear to be different from this series. This result is in good agreement with Figure 13. From the curves

for data sets RNS1201 and RNS1201B, it is obvious that the detector setting (2θ angle) influences the result. Although this effect is not dramatic and the original results in Figure 13 are still recognizable, this 2θ-dependence is a disadvantage of this new method. On the other hand, when comparing Figure 19 with the original pixel analysis in Figure 14, this new method is a vast improvement because the general trends in data quality are visible. This difference is reasonably explained by recalling that SPOTS performs essentially an integration over a reflection, as is done in a more rigorous way during complete data processing, whereas pixel analysis investigates individual pixels in a frame. Upon inspection of the results there is one important question to answer: does the cutoff value used in SPOTS influence the conclusions? In order to address this question, graphs comparable to Figure 19 were made from the same data sets, with various cutoff values. Values of 10, 15 (the program's default), 25, 50, 75, and 100σ were tested. Each σ-value yields essentially the same same results as shown in Figure 19. As an illustration, a plot is shown in Figure 20 for data acquired with a cutoff value of 100σ. When the cutoff value is increased, the curves tend to become more constant, as opposed to decreasing with $1/d^2$. This can be understood by considering the actual meaning of the cutoff value. With a cutoff value of 100σ, the reflection must at least contain one pixel with an intensity 100 times greater than the background. A reflection with several pixels of 50σ, which would still make a valid reflection, is rejected. The average I/σ value of the accepted reflections must therefore increase. Since there are likely more weak reflections at high resolution than at low resolution, this increase should be more pronounced at high resolution.

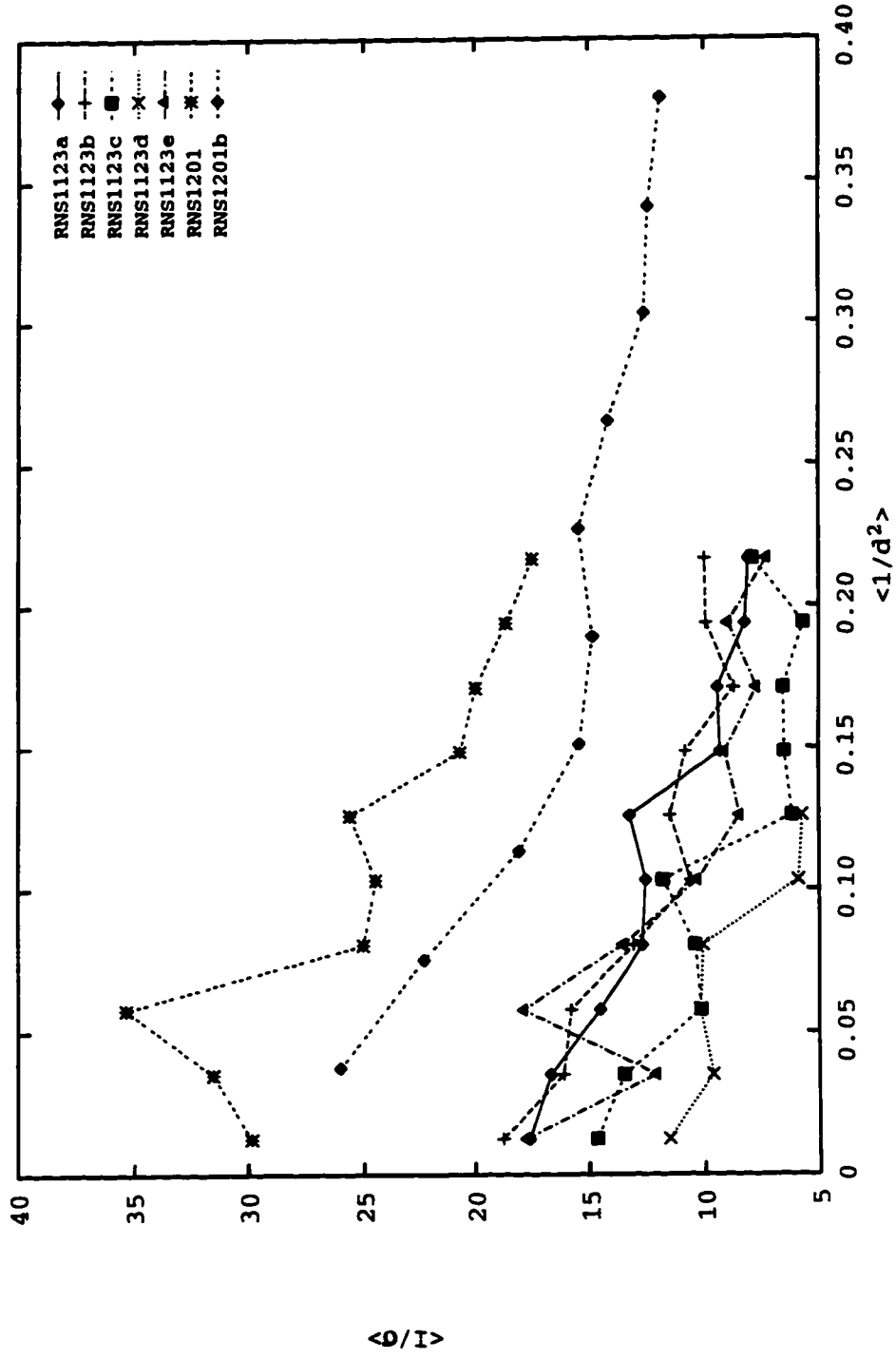

Figure 19: Results for ribonuclease-S obtained with SPOTS.

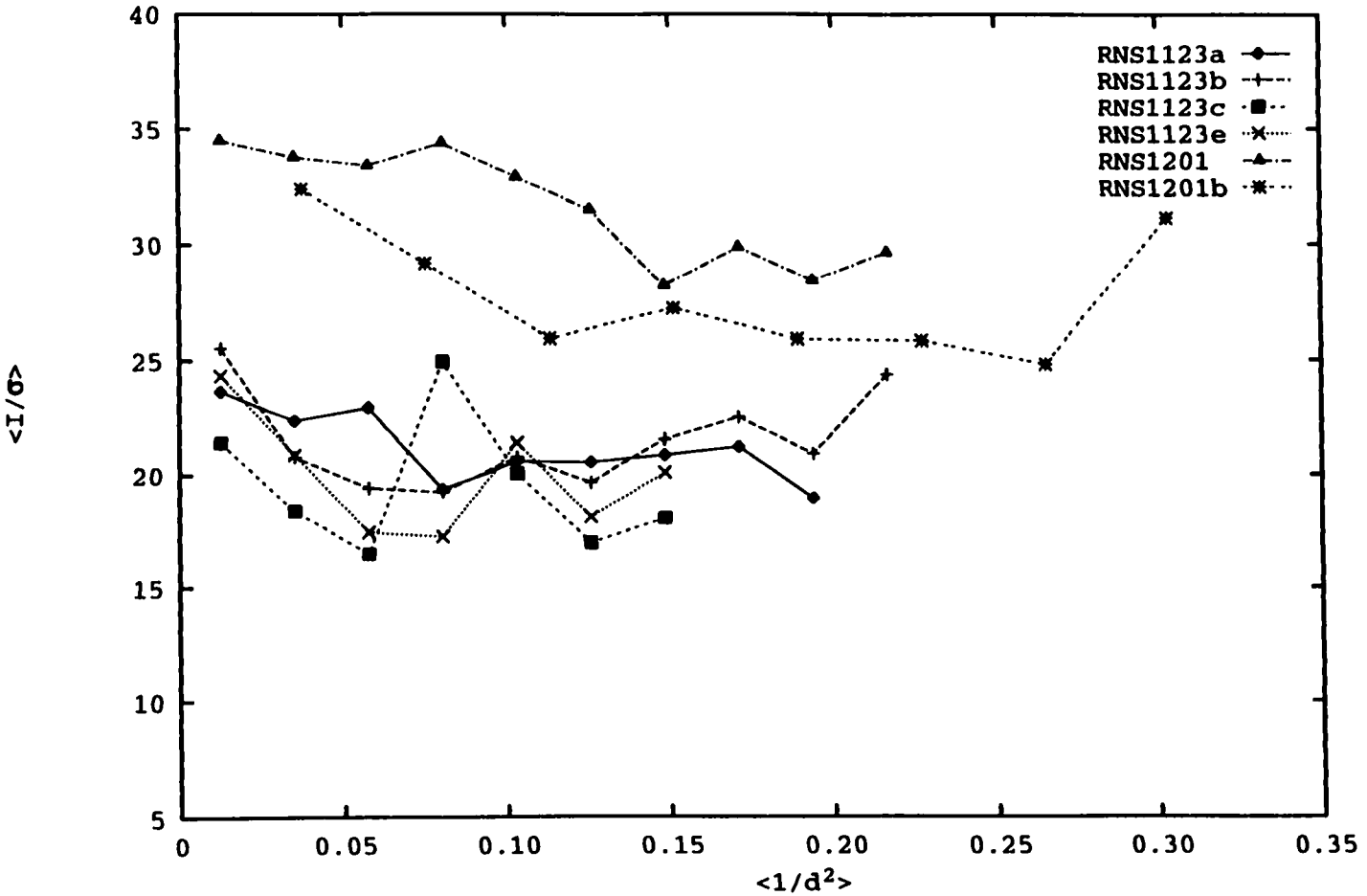

Figure 20: Results for ribonuclease-S obtained with SPOTS, using a cutoff value of 100σ.

2.9.4 Conclusions for the adjusted SPOTS method. The modified SPOTS program represents trends in the data quality better than analysis of individual pixels. This improvement is significant and the modified SPOTS program should be used as opposed to pixel analysis. The results do not appear to be as reliable as those from complete Xengen processing. Particularly the influence of the detector swing angle (2θ) on the I/σ values makes SPOTS less reliable than regular data processing. Thus, SPOTS does not appear to be an improvement over regular data processing, even if much less time is required to obtain data with SPOTS than with full data processing. SPOTS requires minimal user interaction to run.

CHAPTER 3

APPLICATIONS

3.1 Quality Assessment for Lysozyme Crystals Grown on Space Shuttle Flight STS-77.

3.1.1 Introduction. This experiment was performed for various reasons. The first is that is has recently been shown that commercially available lysozyme is not just one (macromolecular) compound and, at the same time, that the impurities in lysozyme affect crystal growth strongly.[19] After initial experiments with ultra-pure lysozyme, a valid question can be asked if this purified lysozyme provides significantly higher quality crystals than do standard preparations. The second point of interest is that standard crystallization conditions[2] for lysozyme call for a pH of approximately 4.7. This pH differs significantly from physiological pH. It is important to determine if, and how much, an enzyme structure changes with pH. This issue leads to the question of whether a structure determined at pH 4.7 gives significant information and structure-function relationships for the same enzyme at physiological pH. Both issues, crystallization behavior and crystal quality, as well as the effects of pH on the structure-function relationship, are the basis for research in this area.

The experiment performed on Space Shuttle flight STS-77 has two practical functions. It serves as a control experiment and as an investigation into the influence of gravity on protein crystal growth. An internal standard is necessary in the event that every protein crystal growth experiment aboard a Space Shuttle fails. If the standard fails

to crystallize as well, this failure reasonably indicates that overall conditions were not suitable for crystal growth. This may be due to sudden movements of the space craft. If most experiments fail, but the standard returns positive, the conditions for the individual experiments were poor, but general conditions for protein crystal growth were met. Lysozyme, because it is relatively easily crystallized, is an adequate standard. The second function of the experiment is to investigate the influence of gravity, or lack thereof, on protein crystal growth.

3.1.2 Materials and methods. Hen egg white lysozyme crystals were grown using the following method. The buffer in which the protein was dissolved consists of 0.1 M Tris-HCl (Tris: Bio-Rad Laboratories, Richmond, CA; HCl: Fisher Scientific), pH = 8.2, and 1 mM NaN_3 (Sigma Chemical Company). The precipitant was either a 9, 10 or 11% (w/v) NaCl (Fisher Scientific) solution in Tris buffer. Lysozyme was made available by Dr. Marc Pusey (MSFC, Huntsville, AL) and was prepared by methods previously described in the literature.[19,22] The hanging drop method[36] was used for crystallization, in hardware especially designed for crystallization in microgravity. The hardware used for this experiment is similar to hardware described in the literature,[9,10,14] but has been changed in order to improve device capacity (number of drops per crystallization unit) and to decrease temperature fluctuations in the device. The commercial vapor diffusion apparatus (CVDA) consists of 32 banks of experiments, each containing four separate experiment chambers. Each chamber has the same function as the wells in Linbro plates (section 2.3.1). A chamber contains a double-barreled syringe which is loaded with protein and precipitant solutions, prior to launch. The bottom of the chamber is fitted with a cylinder of polymer wicking material which holds the reservoir solution. This

solution is the same precipitant as is mixed with the protein solution in all experiments described here. After orbit is attained, a crew member activates the experiments by using a ganging mechanism to extrude the solutions from the syringe barrels to form a protein droplet on the syringe tip. This droplet is then left to equilibrate with the solution in the reservoir. In this particular experiment, drops were made of 20 µl, 50 mg/ml lysozyme (concentration determined by UV absorbance, using an $A^{1\%}_{281.5nm} = 26.4$ according to the literature[45]), and 20 µl precipitant, both in Tris buffer (pH = 8.2). The ground control drops consisted of the exact same solutions with the same concentration and mixing ratios, but were only 15 plus 15 µl in size. In both cases the wells consisted of approximately 1 ml of the appropriate precipitant. The crystals were grown over a period from (Shuttle launch and landing) May 19, 1996, 5:30 AM (CDT) to May 29, 1996, 6:00 AM (CDT), which makes the crystal growth period approximately 10 days. Crystal mounting, data collection procedures and data processing are analogous to the description in section 2.3.1.

3.1.3 Results and discussion. In the case of crystals grown in microgravity, all crystallization conditions (9, 10 and 11% NaCl) provided crystals. On the ground only the condition with 9% NaCl gave diffraction quality crystals. All crystals harvested belong to space group $P4_32_12$. A summary of data is presented in Table 6, for eight crystals grown under microgravity conditions and for three control crystals. For each crystal at least 50 frames of diffraction data were acquired, completely analogous to the description in sections 2.3.1 and 2.3.2. After data processing, both $I/\sigma(I)$ curves and B-factors were determined, as discussed in chapter 2. In Table 6 the entries are first assigned letter F (for flight) or G (for ground control). The size and volume, calculated

by assuming approximately cubic shape of the crystals, is given next to the growth conditions. Finally, two B-factors are given for each crystal. The B_1 values were determined with Wilson plots using all data points, whereas the B_2 values were calculated omitting data points with $(\sin(\theta)/\lambda)^2 < 0.027 Å^{-2}$. The reference point for the Wilson plots was $\Sigma_a f_a^2$ for all atoms in the protein molecule, as derived from the amino acid sequence in entry 1HEL of the Brookhaven protein data bank. The $I/\sigma(I)$ curves are given in Figures 21, 22, and 23. The curves are divided over three graphs for clarity. The labels in the graphs relate to the entries in Table 6 by their letter (F or G) and the volume of the crystal. Thus, F288 refers to the first entry in Table 6.

Table 6: Lysozyme crystals grown on STS-77 at pH = 8.2.

F or G	initial %NaCl	size (mm)	volume *1000 (mm^3)	B1 ($Å^2$)	Corr. coeff.	B_2 ($Å^2$)	Corr. coeff.
F	9	0.9x0.8x0.4	288	8.0	-0.865	4.5	-0.814
F	9	0.9x0.3x0.4	108	8.6	-0.863	4.1	-0.774
F	9	0.9x0.4x0.4	144	8.9	-0.906	6.3	-0.915
F	9	0.6x0.3x0.3	54	7.0	-0.846	5.6	0.840
F	10	0.8x0.8x0.4	256	7.3	-0.827	4.0	-0.809
F	11	0.4x0.4x0.15	24	4.8	-0.759	1.7	-0.655
F	11	0.4x0.3x0.1	12	10.3	-0.934	10.2	-0.930
F	11	0.4x0.3x0.1	12'	3.9	-0.735	1.0	-0.495
G	9	0.9x0.4x0.2	72	6.8	-0.863	3.7	-0.878
G	9	0.85x0.5x0.35	149	5.6	-0.881	2.4	-0.770
G	9	0.7x0.4x0.3	84	-	-	-	-

It can be seen from Table 6 that the volume of the largest flight grown crystal is about twice that of the largest ground grown control, $288*10^{-3}$ mm^3 versus $149*10^{-3}$ mm^3. There is little, if any, change in the B-values, which are very low for all crystals. The crystal labeled F108 in Figure 21 is the best, and the one labeled F256 is the worst, out of the series plotted. This is under the assumption that high $I/\sigma(I)$ values represent good quality. The improvement of F108 over F256 can obviously not be explained with

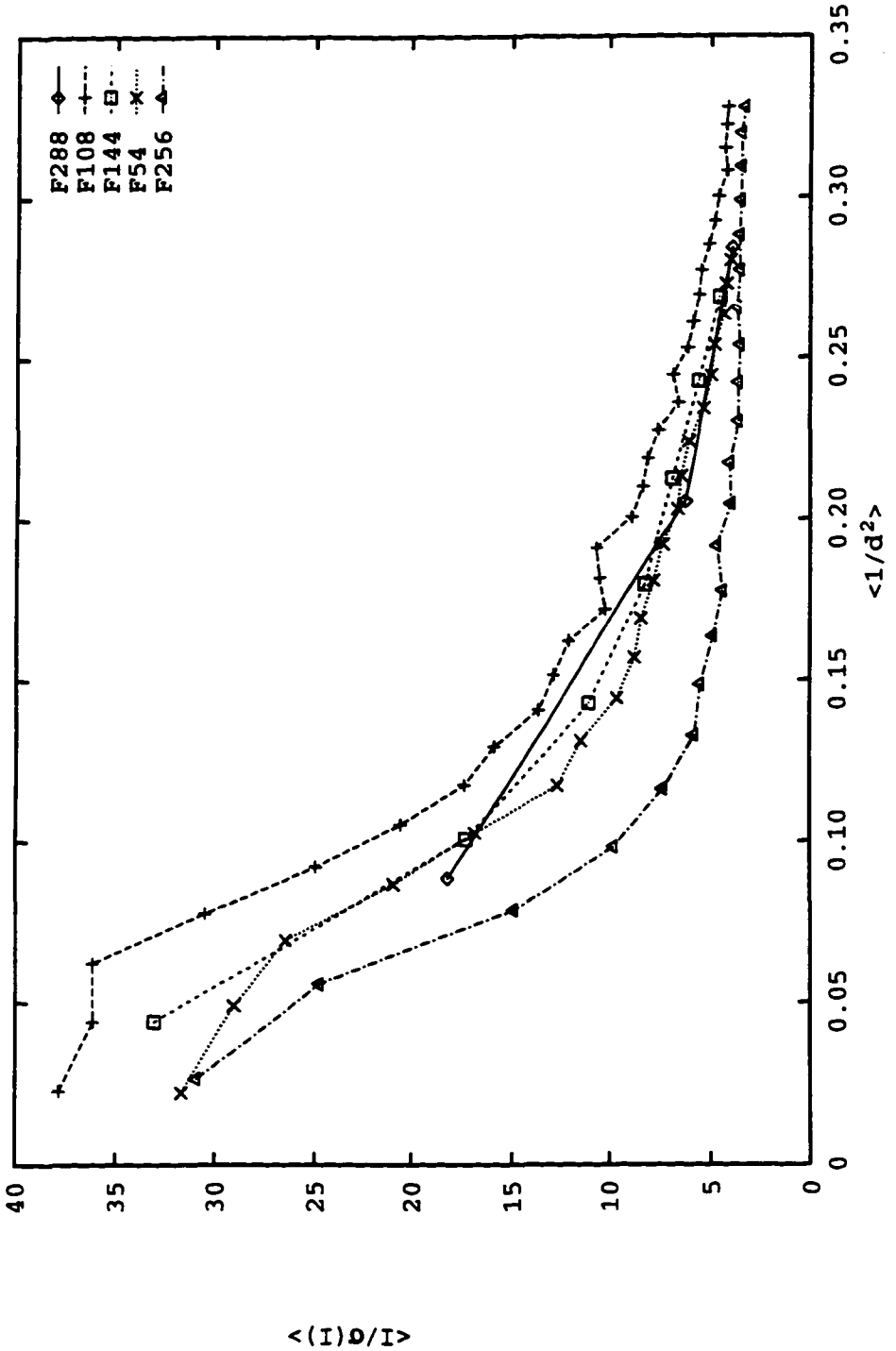

Figure 21: Results for lysozyme on STS-77, part 1.

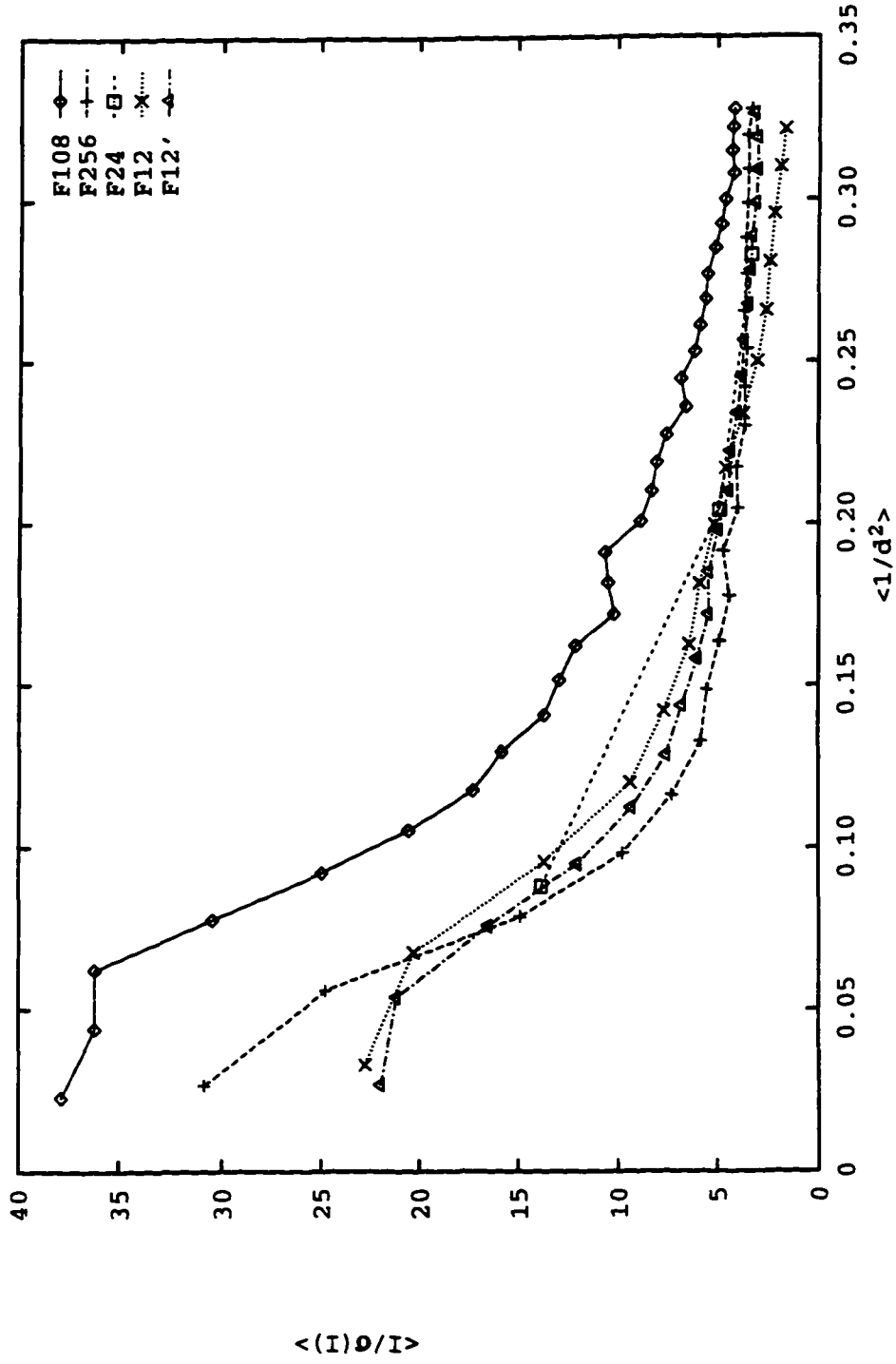

Figure 22: Results for lysozyme on STS-77, part 2.

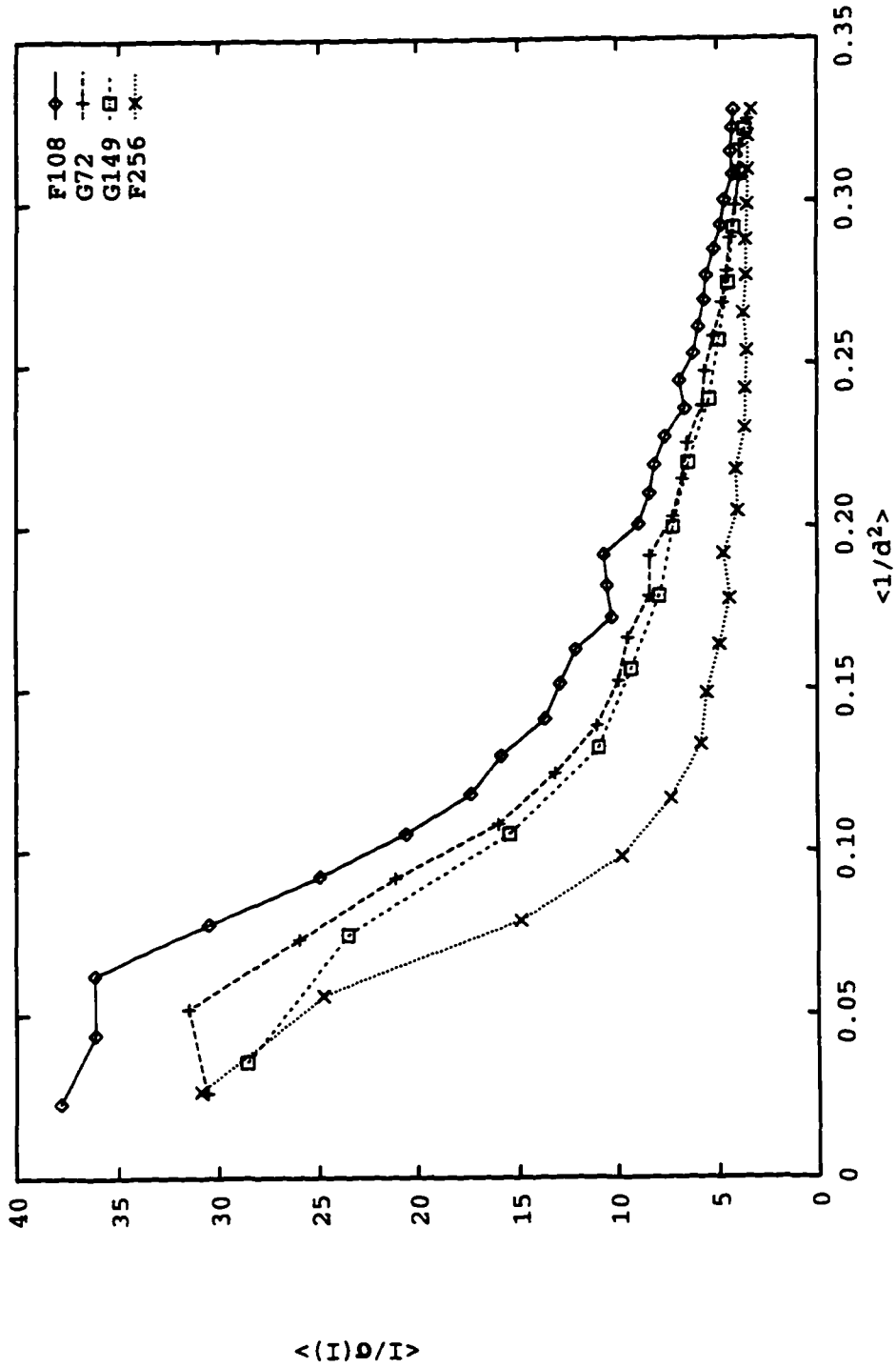

Figure 23: Results for lysozyme on STS-77, part 3.

crystal size. If crystal size were the determining factor, the first crystal with label F288 should be the best. It is possible, since F256 was grown under different conditions than F108, that the use of 9% NaCl as precipitant gives better crystals than the use of 10%. However, this is unlikely since a trend in the crystal quality should then be expected which predicts the use of 11% NaCl to give inferior crystals compared to 10% NaCl. This trend is clearly not observed. The crystals grown using 11% NaCl in Figure 22 are definitely not inferior to F256. Figure 22 further shows one crystal (labeled F12) for which the $I/\sigma(I)$ curve drops off faster than the other curves. This coincides with a larger B-value estimate, as shown in Table 6. This feature becomes very evident at high resolution and emphasizes the importance of a low B-value and the influence on the resolution of the B-value. Data from crystal F12 are of much less value for structure determination than data from other crystals. Finally, in Figure 23, the curves for ground control crystals are shown, together with the curves of the highest and lowest quality crystals grown in microgravity. The two curves for the ground control crystals fall in between the highest and lowest curve, are very similar to each other, and approach the high quality curve closer than the low quality curve. Table 6 shows a third entry for a control crystal, which has no data associated with it. The data obtained from this crystal could not be indexed, most likely because the crystal was not a single crystal. The B-value estimates in Table 6 are accompanied by a correlation coefficient, which is a reasonable indicator for the reliability of the B-value. The estimates for crystals with label F24 and F12′ are different from the other values, but the correlation coefficients indicate that these very low estimates for the temperature factor are less reliable.

There is one observation which indicates that data quality cannot be explained by size. The crystal with label G149 is larger than the crystal with label F108. The data obtained from crystal F108 display higher $I/\sigma(I)$ values than data from G149. This increase clearly cannot be attributed to size.

In this context it is also important to realize that general trends are very important in experiments of this nature. Protein crystals are very fragile and sensitive to dehydration. Mounting a protein crystal, which includes transfer of the crystal from its growth medium to a glass capillary and short exposure to the atmosphere, is probably the step in which most of the variation in crystal quality can be introduced. If crystals are not consistently mounted the same way, using the same procedure, their comparison will not be fair. If one crystal, such as for example crystal F12 from Table 6, seems poorer than the rest, this does not necessarily lead to the conclusion that its growth conditions were poor, but rather that the entire process of crystal growth, crystal preservation and handling was not optimal.

3.1.4 Conclusions for lysozyme experiments aboard STS-77. The crystals grown in microgravity during this flight are all of good quality. The crystals grown in gravity, as control, are also of good quality, and there is no consistent quality change distinguishable between microgravity grown crystals and their control counterparts. The best crystal obtained in this experiment is a crystal labeled F108 and was grown in microgravity. In this experiment the size of the crystals does not appear to be important for the data quality.

3.2 Results for Lysozyme on STS-72

3.2.1 Introduction. In this experiment a standard hen egg white lysozyme crystallization protocol[2] is used at a pH of 4.7. It serves as a control experiment for crystallization conditions aboard the Space Shuttle (similar to the situation for STS-77 described in section 3.1.1). Furthermore, the results can be compared with results at high pH as described in the previous section. This approach allows for an investigation of the influence of pH on crystal quality. Data from one of the crystals grown as control were extensively used in chapter 2 for data verification.

3.2.2 Materials and methods. The buffer used was a 20mM sodium acetate (Fisher Scientific) solution at pH = 4.7. The pH was adjusted with glacial acetic acid (Fisher Scientific). The lysozyme (Seikagaku, lot# E94Z05) concentration was 46 mg/ml, as determined by UV absorbance, using $A^{1\%}_{281.5nm} = 26.4$.[45] Two solutions were used as precipitant: solution A, a 7% (w/v) NaCl (Fisher Scientific) solution in buffer, and solution B, a 8% NaCl solution in buffer. The drops, seven with 7% NaCl and six with 8% NaCl, consisted of 20 µl protein solution and 20 µl precipitant. The wells were filled with approximately1 ml of the appropriate precipitant. The control crystals were grown with the exact same solutions, but in drops consisting of only 15 µl lysozyme solution and 15 µl precipitant. The crystals were grown during Space Shuttle mission STS-72 from (Shuttle launch and landing) January 11, 1996, at 3:41 AM (CST) to January 20, 1996, at 1:41 AM (CST), which makes the crystal growth period approximately 8 days and 22 hours.

3.2.3 Results and discussion. Diffraction quality crystals were obtained from microgravity experiments for conditions with 7% NaCl only. Three crystals were

harvested, two from one syringe (Y-4), one from a second syringe (Y-2). Two crystals were used from the control experiment. Essential information for all crystals is shown in Table 7. The sizes as well as volumes for each crystal are listed. It is clear from the volumes that on average the crystals grown in microgravity are larger than those grown as control. In addition, for each crystal there are two B-values given, with the correlation coefficients for the least squares fit in the Wilson plot. The B_1 value is calculated with all data points included, while the B_2 value only uses data points at resolution higher than 3.16 Å. The B_2 values are likely more reliable, as discussed in section 2.5.1.

Table 7: Lysozyme crystals grown on STS-72 at pH = 4.7.

ID (F or G)	initial % NaCl	size (mm)	volume *1000 (mm^3)	B_1 ($Å^2$)	Corr. corr.	B_2 ($Å^2$)	Corr. coeff.
F (Y-2)	7	1.1x0.6x0.5	330	12.4	-0.949	10.4	-0.954
F (Y-4)	7	0.6x0.3x0.3	54	11.3	-0.938	9.0	-0.947
F (Y-4)	7	0.8x0.55x0.5	220	12.3	-0.949	10.0	-0.969
G	7	0.9x0.6x0.2	96	9.2	-0.868	7.1	-0.878
G	7	0.55x0.4x0.3	66	10.6	-0.929	8.5	-0.943

The $I/\sigma(I)$ curves for each crystal are drawn in Figure 24. The curves indicate two easily recognized features. First, data from crystals grown in microgravity tend to have higher $I/\sigma(I)$ values, as all curves labeled F are above curves labeled G. Second, there is a trend for larger crystals to give higher $I/\sigma(I)$ values. Since the crystals grown under conditions of microgravity (F) are larger, it is not possible to conclude from Figure 24 whether size, microgravity, or both cause higher $I/\sigma(I)$ values. The smallest crystal grown in microgravity, marked F54, is smaller than either of the control crystals, yet data from this crystal show a higher $I/\sigma(I)$ curve. This observation does not support the theory that crystal size is a very important parameter and suggests that other parameters, such as the absence of gravity, may be at least as important.

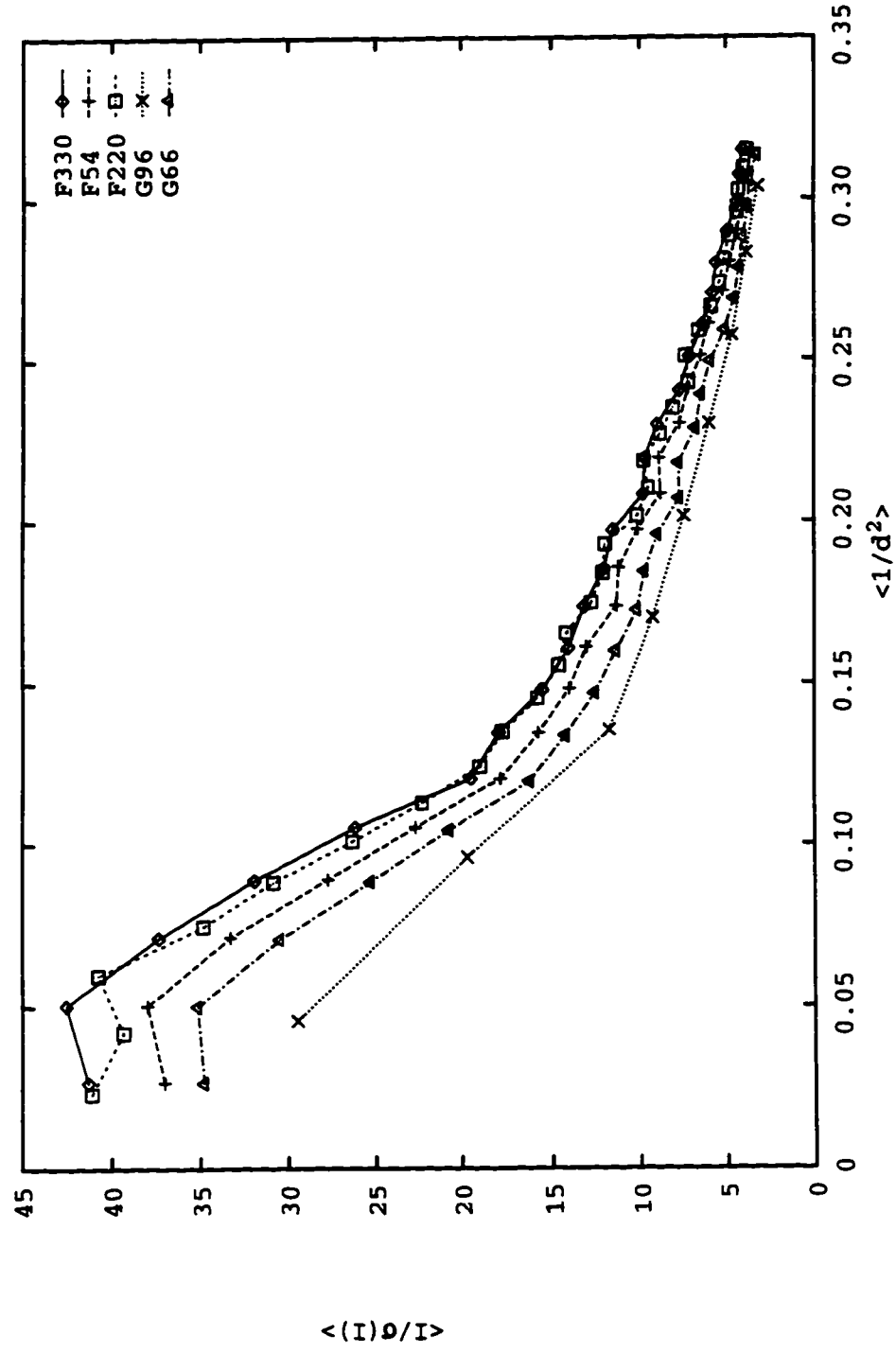

Figure 24: Results for lysozyme on STS-72.

3.3 Comparison of Crystals Grown During Space Shuttle Flights STS-72 and STS-77

When comparing the $I/\sigma(I)$ curves in Figures 21 and 24, the data from STS-72 with lysozyme at pH 4.7 seem slightly stronger. The $I/\sigma(I)$ values for the data from microgravity grown crystals are higher by approximately 5σ at low resolution. At high resolution the best data from both experiments can be interpreted by finding the resolution at which $I/\sigma(I) = 5$. For results at pH 4.7, this point on the curve labeled F330 occurs at a $1/d^2$ value of 0.29. This value translates to a resolution of 1.86Å. The same procedure applied to Figure 21 shows a value for $1/d^2$ of 0.285, using the curve labeled F108. The corresponding resolution is 1.87 Å. The value of $I/\sigma(I) = 5$ was chosen arbitrarily. When the data is compared in this manner, the resolution of the best crystals from each experiment is the same. Since these values are based on intensities I and not on structure factors F, both estimates are lower than the expected final resolution of the structure. This is based on the fact that structure factors are derived from averaged observations, which intrinsically have a higher signal-to-noise ratio.

The B-factor estimates are presented in Table 6, 7, and 8. Table 8 contains the averaged B-factor values for all crystals grown under the appropriate conditions. The pH values and the presence or absence of gravity is a constant for each table entry, but other crystal growth parameters may vary. Comparing the B-factor estimates, two interesting features become obvious. First, the B-factors of the crystals grown in the experiments at pH 8.2 are significantly smaller than B-values of crystals grown at pH 4.7. This is an obvious trend and the explanation may be found in either the effect of pH, or the effect of enzyme purity. The lysozyme used at high pH is of higher quality due to extra purification steps.[19] Even if the relationship between B-factor estimates and final, refined,

B-values for individual atoms in both structures is yet to be determined, these data suggest that crystals grown at pH 8.2 will provide a better defined structure. The decrease in B-factors could be explained by improved internal order in the crystal, or the absence of impurities. Additional experiments are needed to distinguish between pH and enzyme purity as possible causes for lower B-factors. Simultaneous crystal growth experiments can be performed at high and low pH with lysozyme purified to the same extent.

Table 8: Comparison of average B-factor values for lysozyme crystals grown at variable pH.

	Flight B ($Å^2$)	Ground B ($Å^2$)
pH 4.7	9.8	7.8
pH 8.2	4.7	3.1

Second, the B-factors obtained from crystals grown under standard conditions are slightly lower than B-factors from microgravity-grown crystals. It is difficult to conclude from the data presented if the results are significant. It is important that this trend is observed in two independent crystal growth experiments, each at a different pH. This makes the observation more relevant. It certainly is possible that a difference exists between the crystals, because in the absence of convective flows in microgravity, the crystal growth mechanism may be completely different.[7,40] Furthermore, the conditions used for crystal growth in microgravity were optimized for crystal growth on Earth. This implies that the crystals grown in microgravity may not yield the best possible results. Future steps to further understand how B-factors change when crystals are grown in microgravity should include analysis of additional experiments, repetition of the same experiments, and determination of a relationship between B-factor estimates by means of Wilson plots and true B-factors in the final structure.

3.4 Comparison of Crystals Grown With Different Temperature Profiles

3.4.1 Background. As discussed in section 1.2, it is possible to grow lysozyme crystals by means of temperature variation. The temperature profile can be a simple slope, or a more complicated dynamic process, which is guided by observation of nuclei formation. The following section contains the results of such an experiment. All crystals were grown and provided by Rita Gray. X-ray diffraction data were collected and processed by Randy Mann. Data collection took place at 20°C. The data interpretation and results of chapter 2 are of interest to this work because crystal quality can now be judged. For appreciation of the work, part of the procedures and differences in crystal growth methods will be described next. Two different temperature profiles were used to induce and control lysozyme crystal growth. Since the solubility of lysozyme changes with temperature, the supersaturation can be controlled with temperature variation.[8]

One scenario is to slowly decrease the temperature. Results obtained with this method are labeled "slow cool," and the temperature profile is drawn in Figure 25. The variables in this process are start temperature, end temperature, and slope. The starting and ending temperatures are not completely free of choice. Various practical considerations, such as protein denaturation and freezing of solutions limit the range of choice. The slope, however, is an important variable, for which almost any value can be chosen.

The second scenario for using temperature to control crystal growth is called dynamic control. In Figure 25 it is labeled as dynamic. The protein solution is cooled down relatively fast until nuclei are formed, detected by laser light scattering.[30] Subsequently the temperature is raised quickly to a higher value. At this higher temperature the

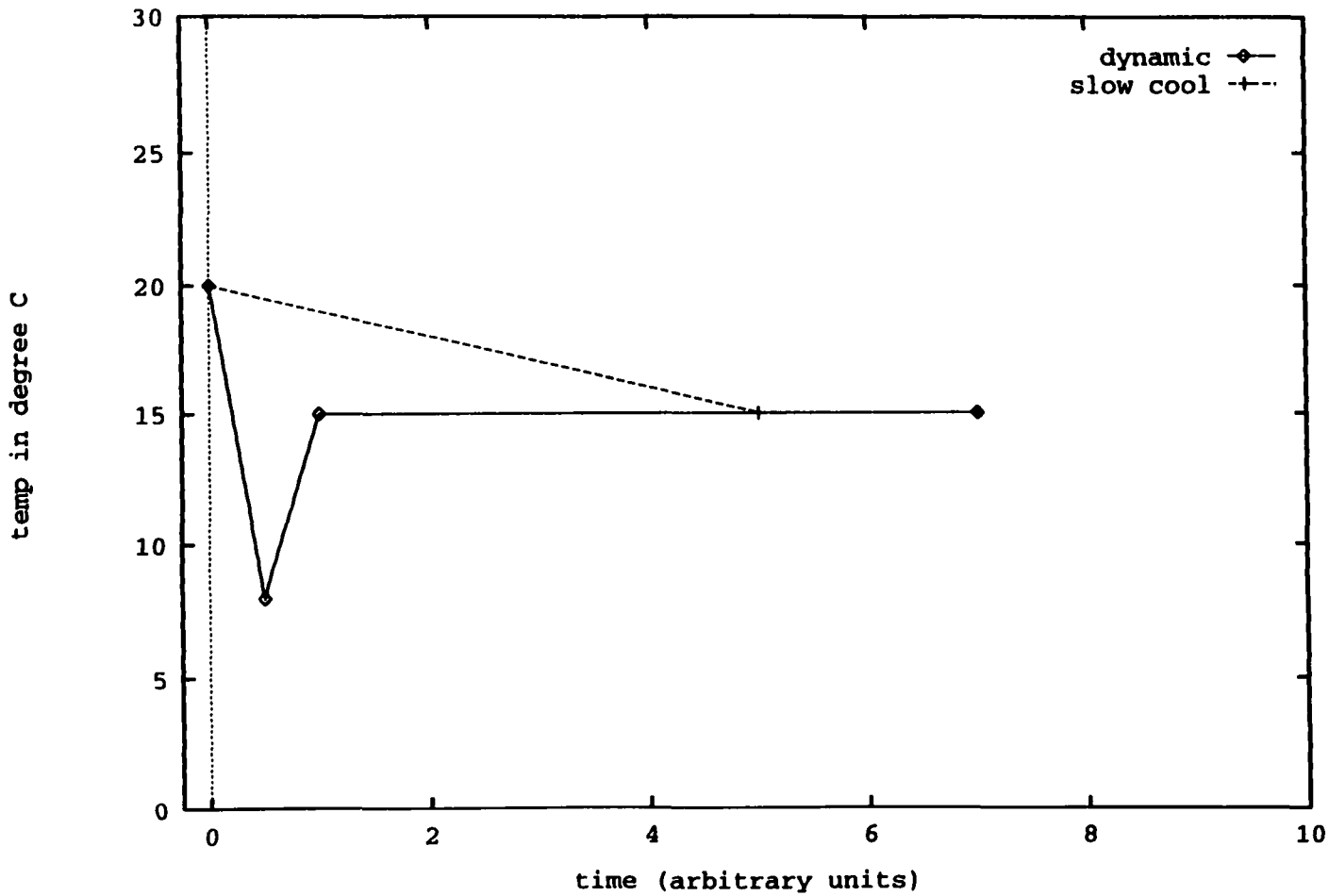

Figure 25: Temperature profiles used for lysozyme crystal growth.

crystals are growing, without (in theory) creation of more nuclei. This process is possible because there are different nucleation and growth regimes (see section 1.2).

3.4.2 Crystal growth experiments. For the dynamic control process, the following procedure was followed. A stock solution of hen egg white lysozyme (Seikagaku, lot#E94Z05) was prepared by dissolving 60 mg of lysozyme in 1 ml of 50 mM sodium acetate (Fisher Scientific) buffer at pH=4.4. Five hundred microliter of this stock solution was mixed with 500 μl of 5% (w/v) NaCl (Fisher Scientific) in 50 mM sodium acetate buffer at pH 4.4. This growth medium was centrifuged for one hour at room temperature. Next, the centrifuged solution was filtered through a 0.2 μm Anotop filter (Fisher Scientific). The growth solution was transferred in aliquots of 200 μl to each of four Starna cells and the cells were incubated at 20°C in the nucleation chambers to establish a laser scattering baseline voltage. The temperature of each of the solutions was then decreased to a nucleation temperature of 8°C at rates of either 0.05°C per minute or 0.025°C per minute. The solution was kept at 8°C until laser light scattering indicated nucleation, or until 24 hours had passed. The temperature was subsequently increased at a rate of at least 0.1°C per minute, to 15°C. The crystals were allowed to grow at this temperature. The crystal growth experiments were performed in hardware especially designed for this purpose. The hardware allows for careful temperature control and observation of the solution by laser light scattering and will be described elsewhere.[24,25]

For the slow cool process, 60 mg lysozyme (Seikagaku, lot#E94Z05) was dissolved in 1 ml of 50 mM sodium acetate (Fisher Scientific) buffered at pH 4.4. A sodium chloride (Fisher Scientific) solution, 5% (w/v) in 50 mM sodium acetate buffer was added one-to-one to the lysozyme stock solution and the resulting solution was spun for one

hour. The supernatant was removed and filtered through a 0.2 μm Anotop filter (Fisher Scientific), and 200 μl aliquots were transferred to each of two Starna cells. The temperature of the growth media was decreased from 20° to15°C at rates of 0.0011°C per minute for chamber #1 (total time elapsed for decrease is three days), 0.00087°C per minute for chamber #2 (decrease requires four days), 0.0007°C per minute for chamber #3 (decrease requires five days), and 0.0006°C per minute for chamber #4 (decrease requires six days).

3.4.3 Data collection and processing. The crystals were mounted in glass capillaries (Charles Supper Co.) with some of the growth medium in the same capillary to maintain a saturated atmosphere. The capillaries were sealed with wax and mounted in a goniometer head on a Siemens multiwire area detector. For the crystals grown with the slow cool method, data collection took place with X-rays at 40 kV and 100 mA (total power 4.0 kW), scanning the ω-axis. Each data set consists of 50 frames of 0.25° per frame. The exposure time is 60 seconds per frame. The camera was positioned in such a way that maximum resolution data were recorded for each crystal, as determined by visual inspection. For crystals grown with the dynamically controlled process, the same mounting and data collection procedures were followed, except that the X-ray beam was produced at 40 kV and 60 mA (2.4 kW of power), or 40 mA (1.6 kW of power). These settings were changed to keep counts on the detector face at a reasonable value.

Data were processed with the Xengen package.[28] The program STATS was used to obtain $I/\sigma(I)$ curves. The only deviation from the recommended procedure in the Xengen manual[27] is the omission of the program REJECT in the sequence. This means that all data are used in the final analysis. The program WILSON (based on CMBISO by

Furey[23]) was used to obtain estimates for the B-factors. The B-factors are absolute, not estimates for Δ B, with as reference point the atomic scattering factors determined by taking the amino acid sequence for lysozyme from entry 1HEL in the Brookhaven protein data bank. The procedures to obtain $I/\sigma(I)$ curves as well as B-factors follow the outline in chapter 2 completely.

3.4.4 Observations. A summary of the crystal data is presented in Table 9. The crystals are ordered in two groups, one group was grown with the slow cool protocol, labeled SC, and the other by dynamic control, labeled D. When the crystals were mounted, only two out of three dimensions were measured, hence the size column in Table 9 gives the size of a crystal face in 10^{-2} mm^2, rather than their volume. The rates, which determine the slopes as depicted in Figure 25, are entered for each crystal in Table 9, as well as an estimate for the B-factors. The estimates are based on Wilson scaling as discussed in section 2.5.1, using data with a cutoff at 3.16Å. An $I/\sigma(I)$ curve is drawn for each crystal in Figure 26. The data indicate that the crystals grown with dynamic control, that is, with a temperature profile first going down to 8°C and then back up to 15°C (see Figure 25) are forming a population of crystals with nearly indistinguishable $I/\sigma(I)$ values.

The $I/\sigma(I)$ curves for the crystals grown with the "slow cool" method, show a well-defined, but broader population of crystals. The size and the growth rate of these crystals could somewhat influence the curves, but at this point it seems that the most prominent feature of the curves is the difference between the set of crystals grown with the "slow cool" technique, and dynamically grown crystals. There is no clear trend that larger crystals give higher $I/\sigma(I)$ values for crystals grown with either technique.

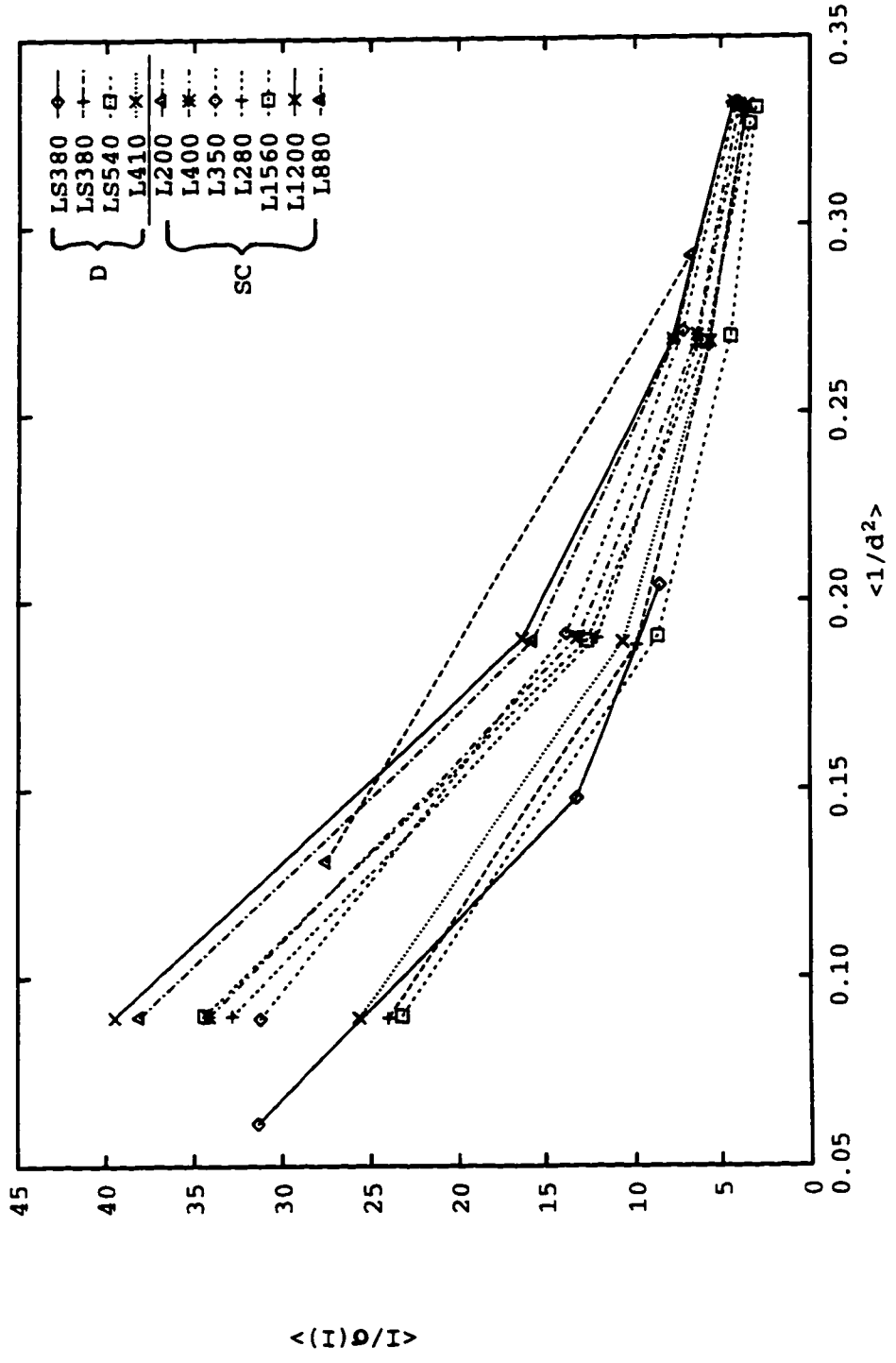

Figure 26: Average $I/\sigma(I)$ values for lysozyme crystals grown with two different temperature profiles.

Table 9: Data for lysozyme crystals obtained by temperature controlled crystal growth.

Method	face size *100 (mm^2)	down rate (°C/min)	B_2 ($Å^2$)	corr. coeff.
SC	0.88	0.0014	12.1	-0.977
SC	1.20	0.003	13.0	-0.963
SC	1.56	0.01	13.5	-0.987
SC	0.28	0.01	11.0	-0.912
SC	0.35	0.0011	13.6	-0.966
SC	0.40	0.002	13.5	-0.988
SC	0.20	0.002	12.9	-0.968
D	0.41	0.025	8.9	-0.935
D	0.54	0.05	10.3	-0.974
D	0.38	0.025	8.1	-0.941
D	0.38	0.025	6.7	-0.809

The B-factor estimates for crystals grown with the dynamically controlled temperature profile are on average significantly smaller than those for crystals, which were grown with the "slow cool" technique. This observation is also clearly visible in Figure 26. The $I/\sigma(I)$ curves for crystals grown with dynamic control decrease at a slower rate as a function of resolution than the curves for crystals grown by cooling solutions slowly.

3.4.5 Discussion. Crystal growth can be thought of as three separate processes: nucleation, actual growth, and cessation. This was discussed in section 1.2. The experiments described here are an effort to manipulate lysozyme solubility with temperature in such a way that nucleation and growth are triggered independently. When a solution is cooled down slowly from 20°C to 15°C, the solubility of lysozyme steadily decreases. The supersaturation increases during this process. Since nucleation is more likely to occur at high supersaturation,[42] nucleation should in theory become more likely as the temperature is lowered. This is the opposite of the desired sequence: first the formation of a few nuclei, then production of a few, large crystals out of these nuclei. Using

dynamically controlled temperature, the nucleation and crystallization events are in the correct order: first the solution is cooled to a low temperature to allow nucleation, then the solution with nuclei is warmed up to an intermediate temperature where the nuclei can grow to form a crystal.

The data presented suggest that there is indeed a difference between crystals grown by using different temperature profiles. The average B-factors for the crystals grown by the "slow cool" method and the dynamically controlled method are 12.8 and 8.5, respectively. A lower B-factor means that the structure is better defined due to more exact repetition of the atoms and molecules in the unit cell, either as a result of decreased vibration, or as a result of more accurate stacking of the unit cells in the crystals. $I/\sigma(I)$ values decrease more slowly as a function of resolution in the case of low B-factors than in the case of high B-factors. Low B-factors therefore may also imply that the structure can be determined at higher resolution.

At this time there are insufficient data available to comment on the influence of the nucleation and growth temperatures, or the slopes in the temperature profile. From Table 9 it appears that B-factor estimates are independent of the slope used when the solution is cooled slowly. With dynamic control the data warrants further research to confirm the favorable influence of the temperature profile.

There is, however, another important issue to be addressed in the experiments as they were performed. The data sets from the crystals grown with dynamic temperature control were collected with the power of the X-ray beam a fraction (60% and 40%) of the power used on the crystals grown with the "slow cool" method. In the next paragraph an

experiment is described to determine whether or not X-ray power is a parameter of importance in the determination of $I/\sigma(I)$ curves and B-factors.

3.4.6 Influence of X-ray power on $I/\sigma(I)$ and B-factors. In order to understand the influence of the power with which X-rays are generated on $I/\sigma(I)$ curves and B-factor estimates, the following experiment was performed:

Three data sets were obtained from the same lysozyme crystal, with the power settings of the X-ray generator at 40kV and 100, 60, and 40 mA, respectively. The crystal used was the crystal labeled F54, as described in sections 3.1.2 and 3.1.3. Each data set consists of 50 frames, each 0.25° in ω, using an exposure time of 60 seconds per frame. For each data set the final structure factors F were calculated in order to determine the B-factors. As discussed in chapter 2, B-factors estimates were calculated with the aid of Wilson plots in two ways, one for all data points (B_1) and one for data points with resolution higher than 3.16 Å (B_2). The $I/\sigma(I)$ values were obtained from program STATS.

The B-factors are listed in Table 10. Figure 27 displays how the average $I/\sigma(I)$ values change as a function of resolution, when the X-ray power varies. There is a trend of decreasing B-factors when the X-ray power decreases. The average $I/\sigma(I)$ values also decrease when power decreases. Lower $I/\sigma(I)$ values are to be expected since a weaker incident beam will cause weaker diffraction and a lower signal-to-noise ratio.

Table 10: B-factor estimates as function of generator power

X-ray power (kW)	B_1 (Å^2)	corr. coeff.	B_2 (Å^2)	corr. coeff.
4.0	7.0	-0.846	5.6	-0.840
2.4	4.3	-0.725	1.9	-0.548
1.6	3.7	-0.652	2.0	-0.473

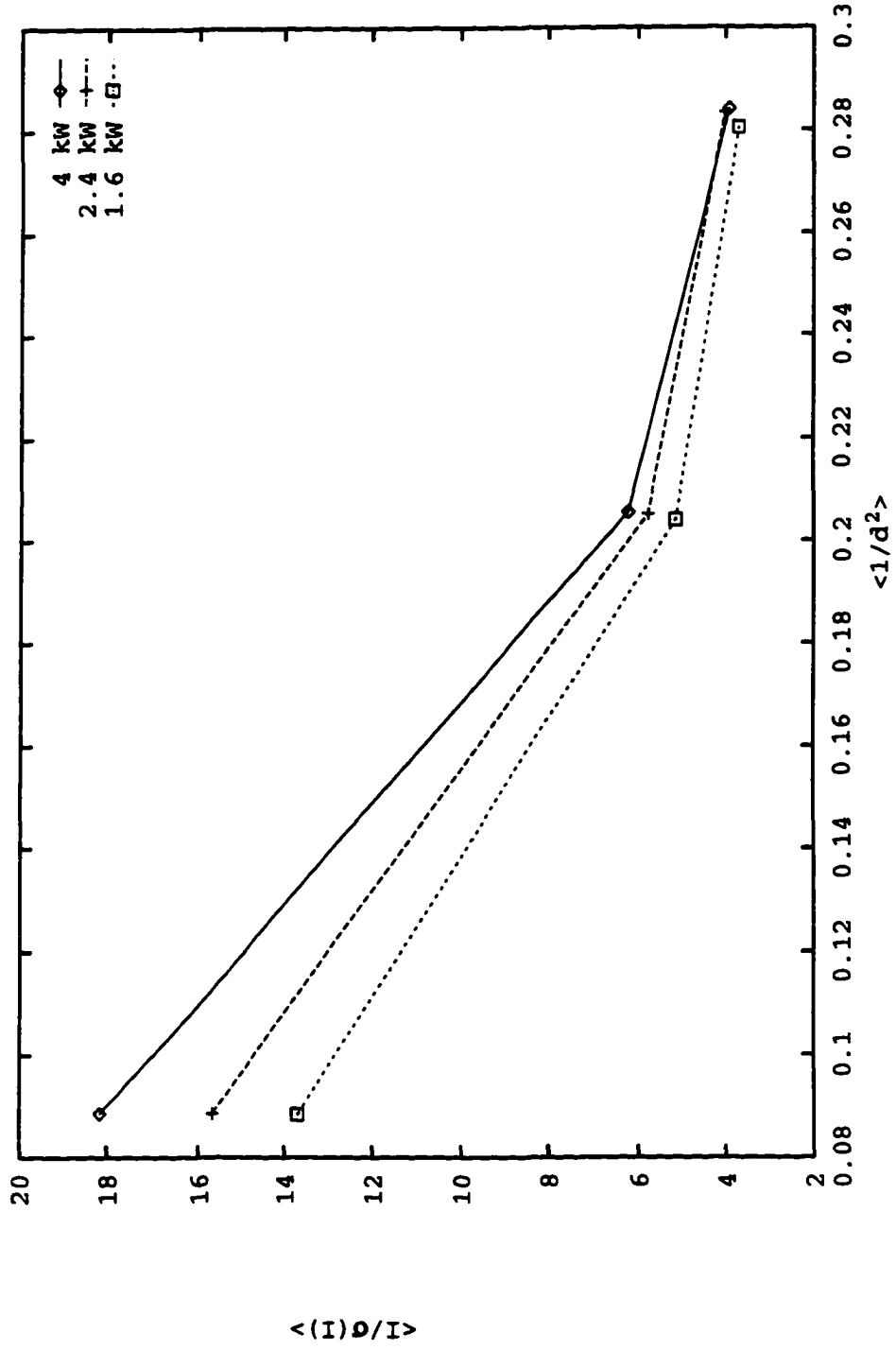

Figure 27: Comparison of average $I/\sigma(I)$ values, obtained with variable X-ray power.

The change in B-factors, however, is very surprising. This change indicates that the signal-to-noise ratio in the data is not equally affected in the entire resolution range. The $I/\sigma(I)$ values at low resolution appear to be more affected than values at high resolution by the change of intensity of incident beam. It is not clear how, or why, this ratio changes.

At the end of data processing the structure factors F are obtained with their individual σ-values. The structure factors are closely associated with averaged intensities Y. Inspection of F and σ_F therefore provides information similar to information obtained through Y and $\sigma(Y)$. Following the conclusions from chapter 2, a comparison is reasonable if the data sets are of equal size and the redundancy is identical for the sets compared. As pointed out above, the size for each data set is 50 frames. The redundancy is exactly equal for each set (1.26). Comparison of F and σ_F values is therefore a reasonable method. The $1/d^2$ values in Table 11 are high resolution cutoff values for each resolution bin. The F and σ_F values are averages for each bin.

The data in Table 11 seem to follow this trend: at low resolution, the F values decrease with decreasing X-ray power, whereas the σ_F values are roughly constant. This causes a decrease in the F/σ_F ratio. At high resolution, however, both F and σ_F values appear to be constant (and hence their ratio). This accounts for the difference in B-factors as observed.

Logically, the next question to investigate is why the intensities change at low resolution. There are two reasonable answers to this question. One possibility is that the detector experienced problems due to the high intensity of the reflections. Reflections at low resolution are more intense than those at high resolution. If there is a problem

Table 11: Values for F and σ_F at various power settings.

	4.0 kW			2.4 kW			1.4 kW		
$1/d^2$	F	σ	F/σ	F	σ	F/σ	F	σ	F/σ
0.05	20.58	0.42	48.5	18.25	0.37	49.0	14.89	0.36	41.5
0.10	16.62	0.44	38.0	14.46	0.39	36.9	12.56	0.41	30.6
0.15	19.03	0.47	40.4	16.37	0.47	35.2	14.35	0.47	30.6
0.20	12.06	0.56	21.5	10.97	0.58	18.8	9.86	0.55	17.9
0.25	9.48	0.81	11.7	8.98	0.78	11.5	8.02	0.73	11.0
0.30	7.94	0.90	8.9	8.33	0.90	9.3	7.58	0.89	8.5

associated with recording high intensities, this would affect low resolution reflections more than those at high resolution. Upon inspection of the area detector log files, it was found that the percentage of late counts on the detector, averaged over 50 frames, was in an acceptable range, and they decreased as the power decreased. These percentages are: 8.58% at 4.0 kW, 6.66% at 2.4 kW, and 4.91% at 1.6 kW (50 frame averages). It is therefore unlikely that there is a counting problem. Detector saturation does not seem to be a problem if it can be judged by these counting statistics.

The second possibility is that there is a fundamental change in diffraction, which is not accounted for in the standard data processing. If protein crystals approach perfection, their structure factors should become smaller because of secondary extinction.[46] This type of extinction arises if a large amount of incident radiation is reflected from the crystallographic plains. This phenomenon has never been observed with protein crystals and is therefore unlikely. If crystals approach perfection, the reflections should also be expected to become very sharp, due to very low mosaicity, which was not particularly evident during data collection.

The problem of decreased B-factor estimates when X-ray power decreases may be resolved by repeating the experiment on a different detector, such as an image plate

system. These detectors work on a different principle. If the changes in data patterns are caused by detector behavior, the same pattern will not be seen on a different detector. Detector saturation may be ruled out in this manner. It may also help to collect much larger data sets and determine if the B-value estimates are reasonably accurate by comparison with refined B-factors. This will rule out completely that the phenomenon is a flaw caused by either poor statistics or inappropriate assumptions in the theory behind B-factor determination.

The general conclusion for this section is that it is very important to keep the X-ray power constant when comparing data from different experiments. When the power changes, the data are clearly affected. There are no direct indicators why the data change with power as they do, but it is a remarkable phenomenon. With respect to the original data in the previous section, it is not possible to determine whether the crystals caused the change in diffraction pattern or if the data collection parameters caused the change. The former possibility indicates that quality change may be a result of different crystal growth methods, whereas the latter indicates that the data collection parameters, specifically the intensity of the incident X-ray beam is the cause. Only a repeat of the entire experiment will provide a more clear answer.

CHAPTER 4

GENERAL CONCLUSIONS AND FUTURE WORK

The conclusions and recommendations from the method development in this work are that both $I/\sigma(I)$ curves, as a function of resolution, and B-factor estimates based on Wilson plots, are good crystal quality descriptors. These conclusions were reached by verification that neither descriptor depends on the data sample size or specific sample contents. Any descriptor that involves averaged intensities Y does depend on data sample size and specifically on redundancy. It is critically important to report data redundancy for each crystal studied when using $Y/\sigma(Y)$ curves. It is preferable to keep the redundancy constant.

These recommendations were used on three sets of experimental data, as a practical test. One pair of experiments shows clearly that it is practical to use $I/\sigma(I)$ curves and B-factor estimates for data quality description. It can be concluded from B-factor estimates, that data obtained from crystals, grown from ultra-pure lysozyme at a pH of 8.2, are of higher quality than data obtained from similar crystals, which were grown from commercially available lysozyme at pH 4.7. With the data currently available, it is impossible to conclude which condition, impurity level or pH, is responsible for the change in data quality. It is unlikely that gravity had as much influence on the crystal quality as the nature of the protein samples or pH. This statement only holds true for this specific comparison.

Curves of $I/\sigma(I)$ as a function of resolution, combined with B-factor estimates provide a good tool to distinguish between different crystals, as shown for lysozyme crystals grown by temperature variation. It was not possible to conclude if the differences observed were due to crystal growth conditions, because of variation in experimental conditions. However, clear differences in data quality were observed by using the recommendations made in this work.

Future work will have to include the establishment of a relationship between B-factor estimates and refined B-factors. This relationship does not need to be an exact correspondence of B-values, but should show a correspondence in the trends of the B-factor estimates with the trends in the refined B-factors. Similarly, it will prove useful to know if there is an advantage in using data with high $I/\sigma(I)$ values, other than resolution extension accompanying high $I/\sigma(I)$ values. A useful experiment which may help answer this question is to refine the same structure twice, once with data consisting of high $I/\sigma(I)$ values and once with data consisting of lower values. Both refinements should be done with data limited to exactly the same resolution.

The most difficult part of research in this field will probably be of a different nature. In an experiment it is not obvious if a crystal is representative of the conditions it was grown in. This is more or less the same statement as saying that it is difficult to ensure that the crystallization conditions are identical for all crystals. This problem was not addressed in this work but will need to be addressed before a link can be made between the product of the process of crystal growth and the conditions used during crystal growth. It is unlikely that extensive statistical analysis, which is measurement of

very many crystals per condition, will provide a solution. Data collection is a laborious and time-consuming process, and only a few crystals can be tested per day at most.

LIST OF REFERENCES

1. Abergel, C.; Nesa, M. P.; Fontecilla-Camps, J. C. *J. Crystal Growth* **1991**, *110*, 11-9.

2. Alderton, G.; Frevold, H. L. *J. Biol. Chem.* **1946**, *164*, 1-5.

3. Asano, K.; Fujita, S.; Senda, T.; Mitsui, Y. *J. Crystal Growth* **1992**, *122*, 323-9.

4. Baker, E. N.; Dodson, G. *J. Mol. Biol.* **1970**, *54*, 605-9.

5. Blundell, T. J.; Johnson, L. N. *Protein Crystallography*; Academic Press: London, 1976.

6. Bourne, Y.; Abergel, C.; Cambillau, C.; Frey, M.; Rouge, P.; Fontecilla-Camps, J. C. *J. Mol. Biol.* **1990**, *214*, 571-84.

7. Bugg, C. E. *J. Crystal Growth* **1986**, *76*, 535-44.

8. Cacioppo, E.; Munson, S.; Pusey, M. L. *J. Crystal Growth* **1991**, *110*, 66-71.

9. DeLucas, L. J.; Bugg, C. E. *METHODS: A Companion to Methods in Enzymology* **1990**, *1*, 105-9.

10. DeLucas, L. J.; Bugg, C. E. *Adv. Space Biol. Med.* **1991**, *1*, 249-78.

11. DeLucas, L. J.; Long, M. M.; Moore, K. M.; Rosenblum, W. M.; Bray, T. L.; Smith, C.; Carson, M.; Narayana, S. V. L.; Carter, D.; Clark, Jr., A. D.; Nanni, R. G.; Ding, J.; Jacobo-Molina, A.; Kamer, G.; Hughes, S. H.; Arnold, E.; Einspahr, H. M.; Clancy, L. L.; Rao, G. S. J.; Cook, P. F.; Harris, B. G.; Munson, S. H.; Finzel, B. C.; McPherson, A.; Weber, P. C.; Lewandowski, F. A.; Nagabhushan, T. L.; Trotta, P. P.; Reichert, P.; Navia, M. A.; Wilson, K. P.; Thomson, J. A.; Richards, R. N.; Bowersox, K. D.; Meade, C. J.; Baker, E. S.; Bishop, S. P.; Dunbar, B. J.; Trinh, E.; Prahl, J.; Sacco, Jr., A.; Bugg, C. E. *J. Crystal Growth* **1994**, *135*, 183-95.

12. DeLucas, L. J.; Smith, C. D.; Carter, D. C.; Snyder, R. S.; McPherson, A.; Koszelak, S.; Bugg, C. E. *J. Crystal Growth* **1991**, *109*, 12-6.

13. DeLucas, L. J.; Smith, C. D.; Smith, H. W.; Vijay-Kumar, S.; Senadhi, S. E.; Ealick, S. E.; Carter, D. C.; Snyder, R. S.; Weber, P. C.; Salemme, F. R.; Ohlendorf, D. H.; Einspahr, H. M.; Clancy, L. L.; Navia, M. A.; McKeever, B. M.; Nagabhushan, T. L.; Nelson, G.; McPherson, A.; Koszelak, S.; Taylor, G.; Stammers, D.; Powell, K.; Darby, G.; Bugg, C. E. *Science* **1989**, *246*, 651-4.

14. DeLucas, L. J.; Smith, C. D.; Smith, W.; Vijay-Kumar, S.; Senadhi, S. E.; Ealick, S. E.; Carter, D. C.; Snyder, R. S.; Weber, P. C.; Salemme, F. R.; Ohlendorf, D. H.; Einspahr, H. M.; Clancy, L. L.; Navia, M. A.; McKeever, B. M.; Nagabhushan, T. L.; Nelson, G.; McPherson, A.; Koszelak, S.; Taylor, G.; Stammers, D; Powell, K.; Darby, G.; Bugg, C. E. *J.Crystal Growth* **1991**, *110*, 302-11.

15. DeLucas, L. J.; Suddath, F. L.; Snyder, R.; Naumann, R.; Broom, M. B.; Pusey, M.; Yost, V.; Herren, B.; Carter, D.; Nelson, B.; Meehan, E. J.; McPherson, A.; Bugg, C. E. *J. Crystal Growth* **1986**, *76*, 681-93.

16. DeMattei, R. C.; Feigelson, R. S. *J. Crystal Growth* **1991**, *110*, 34-40.

17. Drenth, J. *Principles of Protein X-ray Crystallography*; Springer-Verlag: New York, 1995, pp 93-95, 120-2.

18. Dounce, A. L.; Allen, P. *TIBS* **1988**, *13*, 317-20.

19. Ewing, F. L.; Forsythe, E. L.; van der Woerd, M.; Pusey, M. L. *J. Crystal Growth* **1996**, *160*, 389-97.

20. Feher, G. *J. Crystal Growth* **1986**, *86*, 545-6.

21. Fiddis, R. W.; Longman, R. A.; Calvert, P. D. *J. Chem. Soc. Faraday Trans. 1* **1979**, *75*, 2753-61.

22. Forsythe, E.; Ewing, F.; Pusey, M. L.; *Acta Cryst.* **1994**, *D50*, 614-9.

23. Furey, W., Biocrystallography Laboratories, P.O. Box 12055, VA Medical Center, University Drive C, Pittsburgh, PA 15240.

24. Gray, R.; Bray, T.; Kim, L.; DeLucas, L. J.; Wilson, W. W. *J. Crystal Growth*, in preparation.

25. Gray, R.; Kim, L.; DeLucas, L. J.; Wilson, W. W.; Arabshahi, A. *J. Phys.* **1996**, in press.

26. *International Tables for Crystallography*; Hahn, T., Ed.; Volume A; D. Reidel Publishing Company: Dordrecht, 1983.

27. Howard, A. J. *A Guide to Macromolecular X-ray Data Reduction for the Nicolet Area Detector System: The Xengen System, Version 1.3*: Genex Corporation, Gaithersburg, MD, 1988.

28. Howard, A. J.; Gilliland, G. L.; Finzel, B. C.; Poulos, T. L.; Ohlendorf, D. H.; Salemme, F. R. *J. Appl. Cryst.* **1987**, *20*, 383-7.

29. *International Tables for X-ray Crystallography;* Ibers, J. A.; Hamilton, W. C., Eds.; Volume IV; The Kynoch Press: Birmingham, 1974, section 2.2, 71-147.

30. Kam, Z.; Shore, H. B.; Feher, G. *J. Mol. Biol.* **1978**, *123*, 539-55.

31. Littke, W.; John, C. *Science* **1984**, *225*, 203-4.

32. Littke, W.; John, C. *J. Crystal Growth* **1986**, *76*, 663-72.

33. Long, M. M.; DeLucas, L. J.; Smith, C.; Carson, M.; Moore, K.; Harrington, M. D.; Pillion, D. J.; Bishop, S. P.; Rosenblum, W. M.; Naumann, R. J.; Chait, A.; Prahl, J.; Bugg, C. E. *International Journal for Microgravity Research and Applications* **1994**, *7*, 196-202.

34. Lorber, B.; Skouri, M.; Munch, J-P.; Giege, R.; *J. Crystal Growth* **1993**, *128*, 1203-11.

35. McPherson, A. In *Methods in Enzymology*, Wyckoff, H. W.; Hirs, C. H. W.; Timasheff, S. N. Eds; Volume 114, Academic Press: Orlando FL, 1985, 112-20.

36. McPherson, A. In *Methods in Enzymology*, Wyckoff, H. W.; Hirs, C. H. W.; Timasheff, S. N. Eds; Volume 114, Academic Press: Orlando FL, 1985, 120-5.

37. McPherson, A.; Day, J. *Protein Science* **1992**, *1*, 1254-68.

38. Miller, T. Y.; He, X.; Carter, D. C. *J. Crystal Growth* **1992**, *122*, 306-9.

39. Pusey, M. L. *J. Crystal Growth* **1992**, *122*, 1-7.

40. Pusey, M.; Naumann, R. *J. Crystal Growth* **1986**, *76*, 593-9.

41. Pusey, M. L.; Snyder, R. S.; Naumann, R. *J. Biol. Chem.* **1986**, *261*, 6524-9.

42. Rosenberger, F. *J. Crystal Growth* **1986**, *76*, 618-36.

43. Schlichta, P. J. *J. Crystal Growth* **1986**, *76*, 656-62.

44. Smith, G. D.; Ciszak, E.; Pangborn, W. *Protein Science* **1996**, *5*, 1502-11.

45. Sophianopoulos A. J.; Rhodes, C. K.; Holcomb, D. N.; van Holde, K. E. *J. Biol. Chem.* **1962**, *237*, 1107-12.

46. Stout, G. H.; Jensen, L. H. *X-ray Structure Determination*; The Macmillan Company: London, 1968; 195-211.

47. Strong, R. K.; Stoddard, B. L.; Arrott, A.; Farber, G. K. *J. Crystal Growth* **1992**, *119*, 200-14.

48. Vekilov, P. G.; Ataka, M.; Katsura, T. *J. Crystal Growth* **1993**, *130*, 317-20.

49. Wilson, A. J. C. *Nature* **1942**, *150*, 152.

50. Wilson, A. J. C. *Acta Cryst.* **1949**, *2*, 318-21.

51. Wyckoff, H. W.; Hardman, K. D.; Allewell, N. M.; Inagami, T.; Tsernoglou, D.; Johnson, L. N.; Richards, F. M. *J. Biol. Chem.* **1967**, *242*, 3749-53.

APPENDIX A

AWK PROGRAM FOR PROCESSING OF SPOTS RESULTS

```
BEGIN
{
 list = 0;
 y=1;
 totrefl=0;
 d=120;
 x0=254.;
 y0=262.;
 thd=35.;
 nbins=10;
 pi=3.14159265;
 thd=(thd/180)*pi;
 lambda=1.5418;
 maxd=res(511,y0);
 mind=res(0,y0);
 binsize=(maxd-mind)/nbins;
 printf("Detector face from %4.2f to %4.2f Ang\n",
       1/sqrt(mind),1/sqrt(maxd));
 mini=9999.;
 for(binn=1; binn<=10; binn++)
   {
   sumint[binn]=0.;
   nrefl[binn]=0.;
   }
}

function res(x,y)
{
 r=0.2*sqrt((x-x0)*(x-x0)+(y-y0)*(y-y0));
 if(x<x0) r=-r;
 th=atan2(r,d);
 reso=2*sin(.5*(th+thd))/lambda;
 return(reso*reso);
}
```

```
{
 if(list != 0 )
   {
   x=$1;
   y=$2;
   intensig=$8;
   totrefl++;
   bi=(res(x,y)-mind)/(maxd-mind);
   if(bi<0.) bi=0.;
   if(bi>1.) bi=.999;
   binn=1+int(bi*nbins);
   nrefl[binn]++;
   sumint[binn] += intensig;
   if(intensig<mini) mini=intensig;
   }
 if($1=="frame") {list = 1;}
}

END
{
 printf("Found %d reflections\n",totrefl);
 printf("Min value is int/sigma=%10.3f\n",mini);
 printf("There are 10 bins...\n");
 for(binn=1; binn<=10; binn++)
  {
  if(sumint[binn]==0.)
   {
   printf("%7.4f %10.3f\n",mind+binsize*(binn-.5),0.);
   }
   else
   {
   printf("%7.4f %10.3f\n",
        mind+binsize*(binn-.5),sumint[binn]/nrefl[binn]);
   }
  }
 printf("Bins range and number of reflections:\n");
 for(binn=1; binn<=10; binn++)
  {
  printf("Bin %d runs from %5.2f to %5.2f with %d reflections\n",
       binn,1/sqrt(mind+binsize*(binn-1)),
          1/sqrt(mind+binsize*(binn)),nrefl[binn]);
  }
}
```

APPENDIX B

AWK PROGRAM TO DETERMINE THE MOLECULAR FORMULA FROM THE AMINO ACID SEQUENCE

```
BEGIN
{
 C=0;
 N=0;
 S=0;
 O=0;
 H=0;
 total=0;
}

{
 total++;
 if($1=="CYS") {C=C+3; N++; O++; S++; H=H+5; next;}
 if($1=="HIS") {C=C+6; N=N+3; O++; H=H+8; next;}
 if($1=="ILE") {C=C+6; N++; O++; H=H+11; next;}
 if($1=="MET") {C=C+5; N++; O++; S++; H=H+9; next;}
 if($1=="SER") {C=C+3; N++; O=O+2; H=H+5; next;}
 if($1=="VAL") {C=C+5; N++; O++; H=H+9; next;}
 if($1=="ALA") {C=C+3; N++; O++; H=H+5; next;}
 if($1=="GLY") {C=C+2; N++; O++; H=H+3; next;}
 if($1=="LEU") {C=C+6; N++; O++; H=H+11; next;}
 if($1=="PRO") {C=C+5; N=N+2; O++; H=H+8; next;}
 if($1=="THR") {C=C+4; N++; O=O+2; H=H+7; next;}
 if($1=="PHE") {C=C+9; N++; O++; H=H+9; next;}
 if($1=="ARG") {C=C+6; N=N+4; O++; H=H+13; next;}
 if($1=="TYR") {C=C+9; N++; O=O+2; H=H+9; next;}
 if($1=="TRP") {C=C+11; N=N+2; O++; H=H+10; next;}
 if($1=="ASP") {C=C+4; N++; O=O+3; H=H+4; next;}
 if($1=="ASN") {C=C+4; N=N+2; O=O+2; H=H+6; next;}
 if($1=="GLU") {C=C+5; N++; O=O+3; H=H+6; next;}
 if($1=="GLN") {C=C+5; N=N+2; O=O+2; H=H+8; next;}
 if($1=="LYS") {C=C+6; N=N+2; O++; H=H+13; next;}
 printf("Something is strange with the input file\n");
}

END
{
 printf("Found %i amino acids\n",total);
 printf("Molecular formula is: C%iN%iO%iS%iH%i\n",C,N,O,S,H);
}
```

GRADUATE SCHOOL
UNIVERSITY OF ALABAMA AT BIRMINGHAM
DISSERTATION APPROVAL FORM

Name of Candidate Mark Johannes van der Woerd

Major Subject Chemistry

Title of Dissertation Assessment of Protein Crystal Quality by X-ray Diffraction Methods

Dissertation Committee:

, Chairman

Director of Graduate Program

Dean, UAB Graduate School

Date 1/7/97

PS.1439