AN ABSTRACT OF THE THESIS OF

<u>Borries Demeler</u> for the degree of <u>Doctor of Philosophy</u> in <u>Biophysics</u> presented on <u>June 29,</u>
<u>1992.</u>

Title: <u>New Methods for Sedimentation and Diffusion Analysis of Macromolecular Structure</u>

Redacted for Privacy

Abstract approved:_____   _____

Methods are presented to acquire data from analytical ultracentrifugation experiments
by computer using the absorption optical scanning system of the Beckman Model-E
ultracentrifuge. A computer program was written which analyzes sedimentation velocity
experiments by the van Holde - Weischet method and by the second moment method. The van
Holde - Weischet method allows a high resolution analysis of sedimentation velocity data by
eliminating the effects of diffusion on the shape of the moving boundary to provide
sedimentation coefficients for a heterogeneous composition of a sample. The second moment
method obtains the sedimentation coefficient by calculating the second moment point, by which
the sedimentation coefficient is defined. Since it is impractical to manually analyze
sedimentation velocity data by this method, these computer programs make an important
analysis method available to the researcher. Using this computer program, it is now possible
to analyze data to a higher resolution and accuracy than manual analysis of stripchart
recordings would permit. Moreover, the time required for the analysis is greatly reduced. Data
from sedimentation equilibrium experiments are analyzed by χ^2 minimization.

Further, a program was written for the acquisition of data to measure diffusion
coefficients from quasi elastic light scattering experiments with a Langley Ford correlator. The
analysis of autocorrelation spectra from light scattering experiments is performed by the
Levenberg - Marquardt method, which allows fitting of data to nonlinear models. The model
used allows the analysis of multicomponent systems by fitting to a sum of exponentials and a

baseline. Traditional analysis of autocorrelation data by hand was limited to least squares fitting of the data to a linear model of one component without an optimized baseline, often an unrealistic approximation of the system. Analysis of autocorrelation data by nonlinear curve fitting increases both the accuracy and amount of data that can be analyzed.

The development of the PPOL-1 208-n series of plasmids and of the miniplasmid pMX is described. These plasmids were designed to allow studies of *in vitro* transcription and chromatin structure after reconstitution with histones. The plasmids themselves were analyzed by sedimentation and diffusion studies using the computer programs. Sedimentation data is presented which suggests a new method for rapid estimation of S_0 (the sedimentation coefficient at zero concentration) for molecules which show a concentration dependency of the sedimentation coefficient. This is accomplished by linearly extrapolating van Holde - Weischet distributions to zero concentration. Manual analysis of sedimentation velocity experiments to determine nonideality contributions required several experiments, computer analysis can provide this information in a single experiment due to the increased resolution of the method.

Diffusion data for this plasmid DNA is used to demonstrate the feasibility of the multicomponent analysis presented here. Also, sedimentation measurements were carried out on reconstituted chromatin and on the effects of ethidium bromide on reconstituted chromatin. The programs were used to demonstrate significant changes in chromatin structure upon ethidium bromide binding. These changes involved the reduction of S of reconstituted plasmids upon addition of ethidium bromide as well as a reduction of heterogeneity of the sample. The data indicates the possibility of a forced exchange of nucleosomes between plasmids, as well as conformational changes in the chromatin structure.

NEW METHODS FOR

SEDIMENTATION AND DIFFUSION ANALYSIS

OF

MACROMOLECULAR STRUCTURE

by

Borries Demeler

A THESIS

submitted to

Oregon State University

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

Completed June 29, 1992

Commencement June 1993

APPROVED:

Professor of Biochemistry and Biophysics in Charge of Major

Chairman of the Department of Biochemistry and Biophysics

Dean of Graduate School

Typed by: Borries Demeler

Date Thesis is presented: June 29, 1992

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

NEW METHODS FOR SEDIMENTATION AND DIFFUSION ANALYSIS OF MACROMOLECULAR STRUCTURE

1. INTRODUCTION

Analytical ultracentrifugation and quasi elastic light scattering experiments have become widely accepted as well proven methods for the hydrodynamic analysis of macromolecular assemblies. These methods for the measurement of sedimentation and diffusion can be used to study molecular weight, composition and molecular dimensions [1]. However, many of the possible applications of these methods have previously been hindered by the necessity for laborious or even impractical calculations.

With computers available to assist this task, several advantages exist:

(A) Because of automation and time savings, the amount and quality of data which can be collected and analyzed is greatly increased.

(B) The accuracy of calculation is not influenced by the subjectivity of the operator.

(C) Sophisticated data analysis methods can be implemented which otherwise would not be available for manual analysis. Since most manual data analysis methods rely on linear models analyzed by least squares fitting, they are often limited to single component analysis.

(D) Computer analysis can enhance the information derived from an experiment by both increasing the resolution of data and allowing the fitting of data to nonlinear models and multicomponent systems.

In this study computer methods are presented for the acquisition and analysis of data from sedimentation experiments from the Beckman Model-E analytical ultracentrifuge and diffusion experiments from the Langley Ford correlator.

2. NEW METHODS FOR COMPUTER ANALYSIS OF

SEDIMENTATION AND DIFFUSION DATA

2.1 Analytical ultracentrifugation

2.1.1 Overview

Analytical ultracentrifugation has played an important role in the characterization of structural features of macromolecular assemblies. While sedimentation velocity experiments reveal information about size, shape and composition of a sample, sedimentation equilibrium allows quite accurate determination of molecular weights.

The physical basis for sedimentation velocity experiments are three forces acting on a solute particle in solution: Centrifugal, buoyant and frictional drag. These forces will balance out once the solute particle has obtained a constant sedimentation velocity in the centrifuge cell:

(1)
$$F_{centrifugal} + F_{buoyant} + F_{frictional} = 0$$

on a molar basis this yields:

(2)
$$\frac{M(1 - \bar{v}\rho)}{Nf} = \frac{v}{\omega^2 r} = s$$

where M is the molecular weight of the solute particle, $\bar{v}$ its partial specific volume, ρ the density of the solution, N Avogadro's number, f the frictional coefficient of the solute particle, v the velocity of the sedimenting solute particle, $\omega^2 r$ the centrifugal field strength at radius r and s the sedimentation coefficient (Svedberg units, measured in seconds $* 10^{-13}$) [2]. Since the velocity of sedimentation can be expressed as dr/dt, Equation (2) can be rewritten as:

$$(3) \qquad \frac{dr}{dt} = r\omega^2 s$$

which upon integration yields:

$$(4) \qquad ln\frac{r(t)}{r(t_0)} = \omega^2 s(t - t_0)$$

In the centrifuge cell, all solute molecules are initially uniformly distributed throughout the cell. When a centrifugal field is applied, the solute molecules start sedimenting to the bottom of the cell and the region at the top of the cell becomes depleted of solute molecules. A moving boundary is formed and if no diffusion of solute molecules could occur, this boundary would remain infinitely sharp. However, since a finite diffusion of solute molecules exists, the boundary spreads with time and a concentration gradient is formed in the cell. If solute molecules have a large diffusion coefficient, the boundary will spread quickly, while large, extended molecules will diffuse slowly and the boundary will remain quite sharp.

In sedimentation equilibrium experiments, the centrifugal field is much smaller and is chosen such that the net flow of the solute particles ceases and sedimentation will be in equilibrium with diffusion at every point within the cell. A gradient is formed at equilibrium whose analysis allows the calculation of the molecular weight of the solute particle [3].

2.1.1.1 Experimental setup
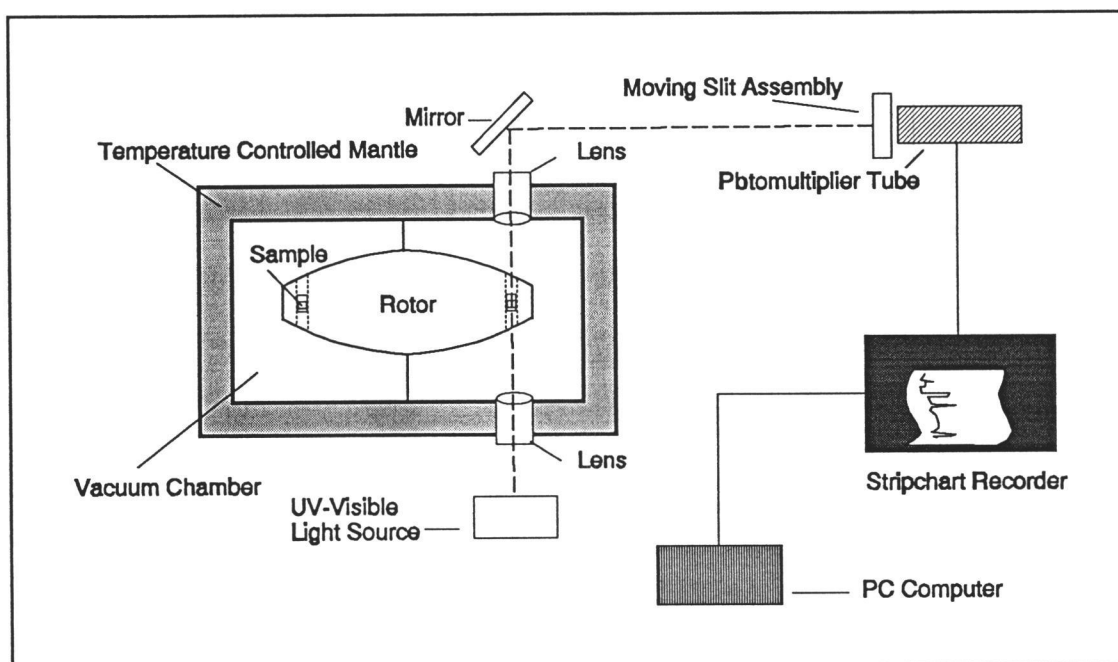
Figure 2.1.1: Schematic view of analytical ultracentrifuge. Absorption within the cell is measured and translated into voltages in the stripchart recorder. In the computer, voltages are digitized by a data acquisition board.

The concentration gradient which forms during sedimentation can be measured by one of two techniques: Interference optics or absorption scanning optics. Interference optics requires laborious measurements of interference fringes to obtain the data, and extremely careful alignment of the optical system. In this study, the much more versatile absorption optical scanner was interfaced with a computer.

The absorption optical system in the Beckman Model-E analytical ultracentrifuge provides a trace of the concentration gradient (on a stripchart recorder) within the centrifuge cell at any given time during the sedimentation process. Light pulses are measured by synchronizing the rotor speed with the sampling of 2 or 4 cells in the rotor. One of these cells is a reference cell providing reference marks of known radial distance to calibrate the radial distance on the absorption trace. The trace is produced by driving a movable slit across the entire width of he cell. Absorption data can be calibrated to optical densities (OD) by a
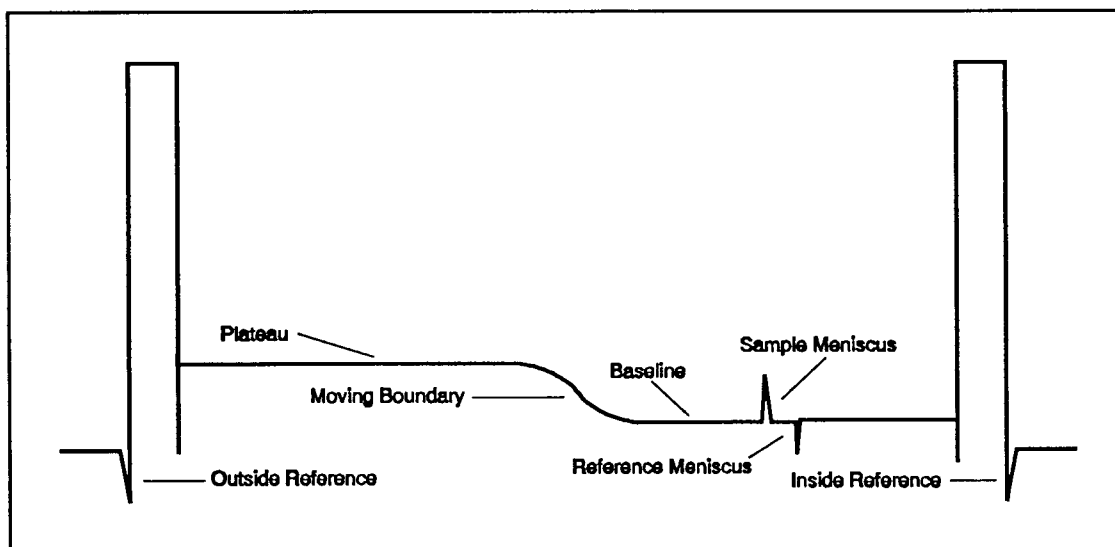
Figure 2.1.2: Schematic rendition of absorption trace of a sedimentation velocity experiment as produced by the optical scanning electronics.

procedure utilizing stairsteps. Stairsteps are discrete absorption values produced electronically in the Beckman Model-E ultracentrifuge. Absorption values in steps of 0.1 OD are provided for calibration of absorption data acquired by the Model-E. The trace can be analyzed by several methods and can provide information about different characteristics of the system.

2.1.1.2 Sedimentation velocity experiments

A) Midpoint method

For a single component system, the midpoint method assumes the species sedimenting at the steepest slope in the sedimentation boundary to be representative of the sedimentation velocity of the solute particle. It is assumed that the spreading of the boundary due to diffusion is symmetric about this point and that it therefore represents the true sedimentation coefficient of the solute particle.

Sedimentation coefficients are obtained by time correcting each scan and determining the sedimentation velocity of the species at the point of steepest slope (called midpoint, although not necessarily the midpoint) in the boundary of the concentration gradient. A graph

of ln($r_{midpoint}/r_{meniscus}$) versus the corrected time of each scan is constructed, and the slope of this graph is approximately equal to $\omega^2 S$. This method will provide an approximate sedimentation coefficient for a single component system. If more than one component is present, the midpoint provides only an unreliable average of all components present in the system. The midpoint method is only chosen for its convenience in manually analyzing sedimentation data [4].

B) Second moment method

For a single component system, the sedimentation coefficient S is defined by the velocity of motion of the second moment point in the absorption gradient, using equation (4). If more than one component is present, the second moment method provides the weight average sedimentation coefficient of all components in the system. It is obtained by integrating the following equation:

$$(5) \qquad \|\bar{r}\|^2 = \frac{\int r^2 * \frac{\delta A}{\delta r}\, dr}{\int \frac{\delta A}{\delta r}\, dr}, \quad (r = \textit{radial position}, A = \textit{absorbance})$$

If there were no diffusion, the midpoint would be the same point as the second moment point. To determine the sedimentation coefficient from second moment points in a number of scans, a graph of ln($r_{second\ moment\ point}/r_{meniscus}$) versus the corrected time of each scan can be constructed, and the slope of this graph is equal to $\omega^2 S$. Knowing the angular velocity ω the sedimentation coefficient S is readily available. This method will provide an the exact sedimentation coefficient for a single component system. However, if the system is heterogeneous, individual components cannot be distinguished by this method.

C) van Holde - Weischet method

The van Holde - Weischet method improves substantially upon the midpoint method by taking into account heterogeneity and diffusion. A system of two or more components may result in two or more distinctly visible boundaries in the concentration gradient, but if diffusion is large, the distinct boundaries may become obliterated due to diffusion. An additional complication is the dependence of sedimentation on concentration which artificially sharpens the boundary, making analytical corrections for diffusion even more difficult.

If the sedimentation pattern could be corrected for diffusion, even cases of high polydispersity could be resolved. While the spreading of the boundary due to polydispersity is proportional to the first power of time, the spreading of the boundary due to diffusion is proportional to the square root of time. Therefore, at infinite time, spreading of the boundary due to polydispersity will dominate the diffusion effect. By plotting apparent sedimentation coefficients (measured at different points in the boundary) versus the inverse square root of time, the van Holde - Weischet method provides a method to make the correction for diffusion [5]. In order to solve the problem of diffusion correction, the distance between absorbance at the baseline and absorbance at the plateau is divided into n horizontally equally spaced intervals for each scan, and the apparent sedimentation coefficient S^* for each subdivision is calculated from the radial position of each intersection of the horizontal lines with the boundary. S^* values for each division are plotted against the inverse square root of time of each scan. S^* values for the nth division of each scan are fitted to a straight line by linear least squares and extrapolated to infinite time ($1/t^{1/2} = 0$). The intersection of this line with the ordinate ($t = \infty$) indicates the true sedimentation coefficient of the species at the nth division in the boundary.

A distribution plot can then be constructed, which shows the percentage of divisions of the boundary sedimenting with a given sedimentation coefficient or less. Relative amounts of components can be approximated with this plot, although the Johnston - Ogston effect may distort the true ratios under some circumstances [6].

2.1.1.3 Sedimentation equilibrium experiments

A) Single component analysis

When a solute is centrifuged to equilibrium, all flow of solute particles will vanish. Sedimentation and diffusion of the solute will balance out, and no net transport in the centrifuge cell will occur. This situation can be represented by the flow equation, with the flow equal to zero [7]:

$$(6) \qquad L[\omega^2 r M(1 - \overline{v}\rho) - \frac{RT}{C}\frac{\delta C}{\delta r}] = 0, \quad L = \frac{C}{Nf}$$

where L is the transport coefficient, ω the radial velocity, r the radius, M the molecular weight, $\overline{v}$ the partial specific volume, ρ the density of the solvent, T the absolute temperature, R the gas constant and C the concentration. Rearranging, integrating from the meniscus position a to some point r in the cell and expressing logarithms as exponentials yields [8]:

$$(7) \qquad C(r) = C(a) * \exp[\frac{\omega^2 M(1 - \overline{v}\rho)(r^2 - a^2)}{2RT}]$$

The single component analysis is performed by subtracting a measured or calculated baseline from each point in the absorption trace and plotting ln(A) versus r^2. If only a single component is present and the baseline is properly estimated, a linear graph can be constructed by least squares fitting whose slope is proportional to the molecular weight of the species present.

B) Multiple component analysis

An alternative is to fit the absorption trace to a nonlinear model of several components by performing a Levenberg-Marquardt χ^2 minimization [9]. The appropriate model to fit is:

$$(8) \qquad A(a_i, \ b_i, \ c, \ r) = \sum_{i=1}^{n} [a_i * e^{b_i * r^2}] + c$$

where a_i, b_i and c are adjustable parameters and a_i corresponds to the amount of component I, b_i is proportional to the molecular weight of component I, c is the baseline, n the maximal component, r the radial distance from the center of the rotor and A is the absorbance at point r [10]. In practice, the resolution of data is generally good enough to allow fitting to a system of two components. This model is of particular interest in the study of monomer - dimer interactions and in association studies of biological macromolecules as found in DNA/protein interactions.

2.1.1.4 Computer implementation for the analysis of sedimentation data

While manual analysis of stripchart paper is laborious and time consuming, it also is prone to introduce inaccuracies and subjectivity into the result. It is clear that computer analysis of sedimentation data by implementing above mentioned analysis methods in a robust algorithm can improve on manual methods tremendously, both in terms of accuracy as well as in time expense. To implement such analysis on the computer in a user friendly environment has been a major aim in this study.

A program (UltraScan) was developed to (1) acquire ultracentrifugation data and (2) analyze sedimentation velocity and sedimentation equilibrium data. The program consists of several independent modules, which are driven by the main program, US.EXE. A complete description of the program is included in Appendix A. The program includes many features designed to automate as much as possible the process of data acquisition, data editing and data analysis. Following is an overview of program structure and module arrangement:
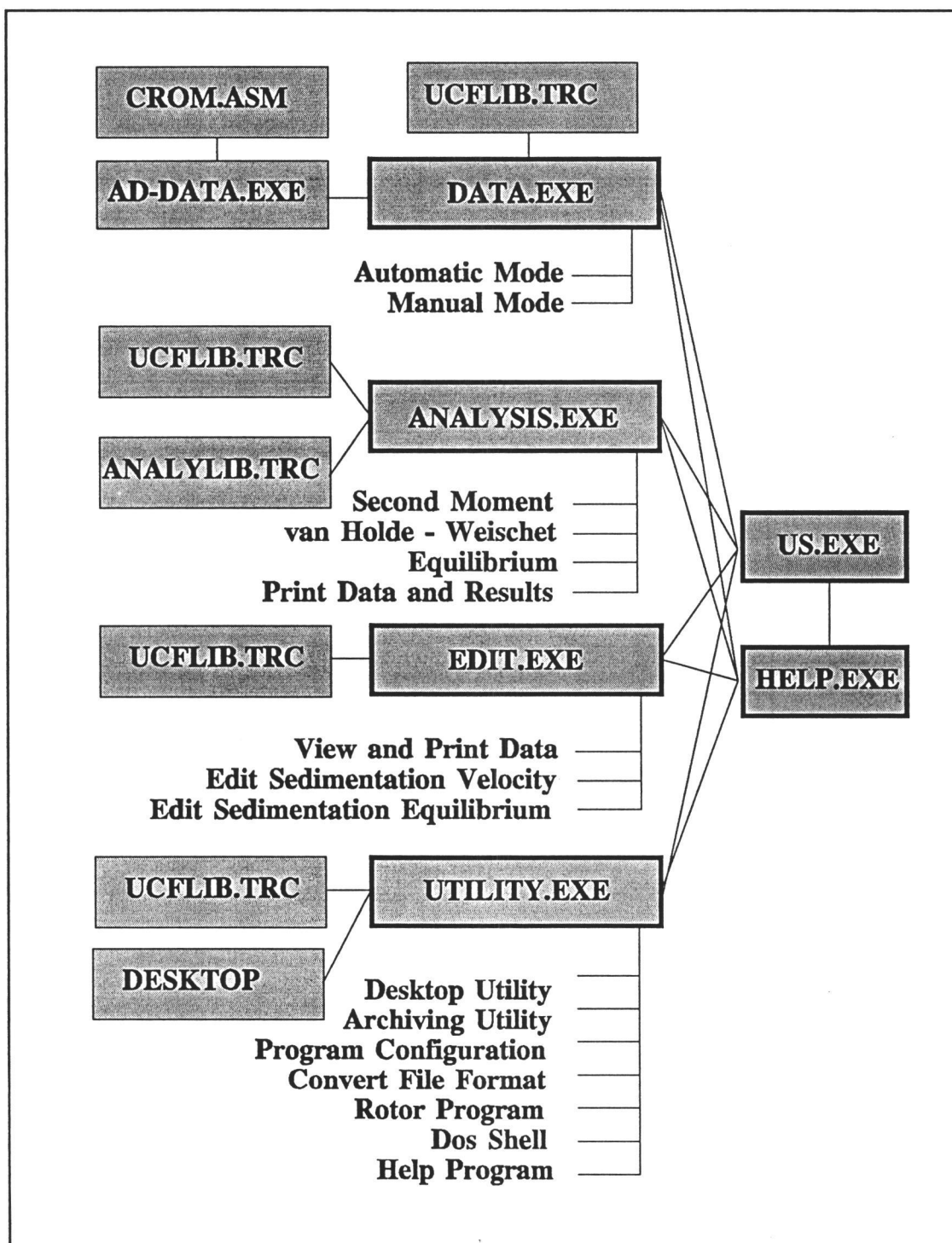
Figure 2.1.3: Program structure for UltraScan. Shown are individual modules (thick borders) and their dependencies (thin borders). Lines represent pathways of accessing subprograms.

2.1.2 Data acquisition

Since the Beckman Model-E analytical ultracentrifuge is not equipped with a computer interface, an interface was designed to convert voltages from the stripchart recorder into frequencies useful for computer analysis. The interface was designed by Carl Baker Industries, Corvallis, Oregon (see also Appendix A, page A-80, A-81). The voltage to frequency conversion was accomplished by a Metrabyte CROM-1 AT data acquisition board (Metrabyte Corporation).

The interface program consisted of three modules. The first module was written in Microsoft assembler and was supplied by Metrabyte Corporation to communicate directly with the data acquisition board. The second module was written in Microsoft Quick Basic and provided the calling routine for the first module. This module (also provided with the Metrabyte CROM-1 AT board) was modified to suit the particular needs for interaction with the UltraScan data acquisition program.

Data acquisition was started by a trigger signal from the Model-E. The Model-E was set to cycling mode with the desired scan interval, and each new scan triggered acquisition of data into a new datafile. The program is able to discriminate between up to six different cells, provided they are scanned consecutively. The time of each scan is read from the computer clock (accurate to at least within 1/10 of a second), to avoid inaccuracies resulting from the mechanical timer on the Model-E.

Each newly acquired scan file contains following information: Run identification, cell number, scan number, time of scan, number of datapoints per second (data acquisition rate), and absorbance data for the entire cell and a little bit outside the cell (for determination of reference positions). The acquisition rate can be varied to allow different resolutions of data. The maximum resolution is determined by scan duration and the number of datapoints acquired per second, the limit is set at 10,000 datapoints (for example, if a resolution of 100 datapoints per second is desired, the scan duration cannot exceed 100 seconds). If desired, stairsteps can

be acquired as well and used for calibration of the absorbance data. Sample data for cytochrome C oxidase as acquired by the Model-E interface is shown in Figure 2.1.4. Conditions for this experiment were as follows: Temperature: 19.8°C, rotor speed: 38,750 rpm, wavelength: 422 nm, concentration of protein: 1.1 mg/ml, buffer: 20 mM Tris-HCl pH 8.1, 1 mM EDTA, 10% Triton X, 90 mM NaCl. The density corrections for this buffer is 1.004, the viscosity relative to water is 1.143 [11].
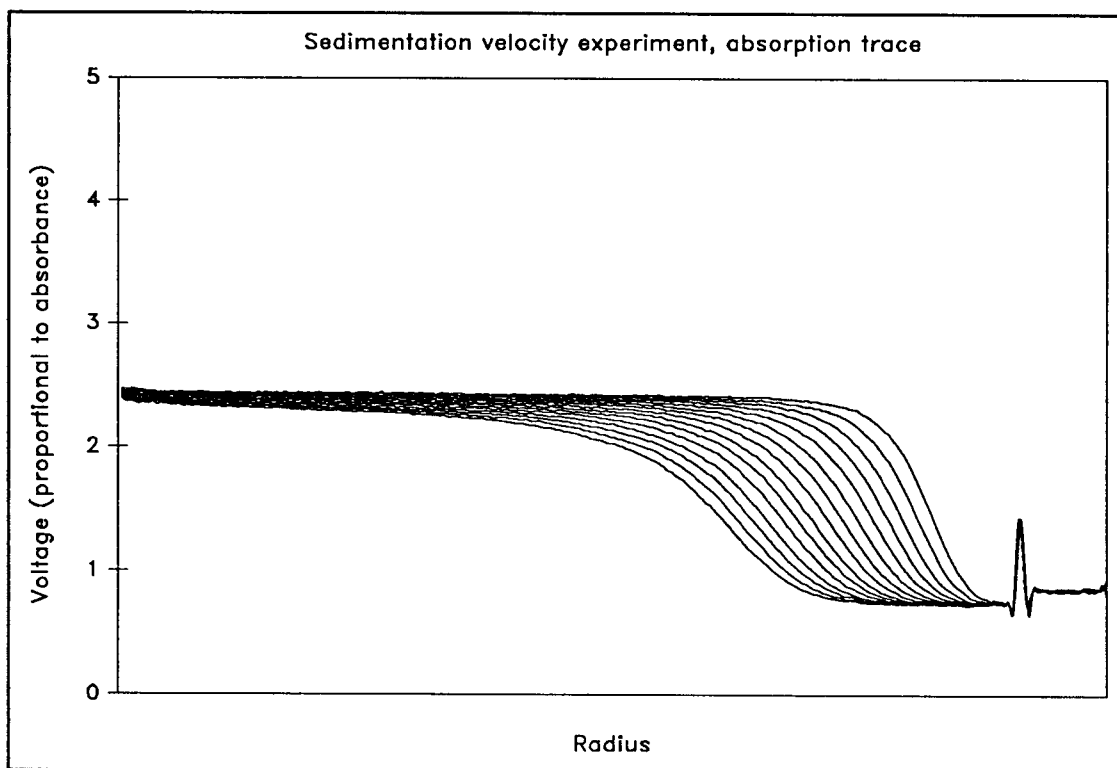
Figure 2.1.4: Sedimentation velocity data for cytochrome C oxidase, acquired by Model-E interface. (Data courtesy of Dr. Neal C. Robinson and Dr. Jeff Hansen, University of Texas, San Antonio)

2.1.3 Analysis methods

2.1.3.1 Editing data

The greatest challenge in designing a program capable of analyzing data is to find an algorithm robust enough to succeed for any imaginable data configuration. Much testing and

debugging work is required to adapt an algorithm in such a way that it can cope with even the rarest case, without sacrificing accuracy and the scientific method. Raw data is hardly ever in a format useful for scientific analysis. An editing program was developed to organize raw data into a standard format useful for all analysis methods applied. Two analysis methods were implemented for sedimentation velocity experiments: van Holde - Weischet and second moment. The requirements for sedimentation equilibrium experiments are different, therefore, a different editing routine was used. The requirements for analysis of data from sedimentation velocity experiments are:

(A)   All scans have to be aligned against each other.

(B)   Absorbance data needs to be aligned with the proper radial position.

(C)   The radial positions and the cell length correspondent have to be calculated in units of time.

(D)   The meniscus position in each scan has to be defined with reference to a fixed point (the inside reference).

(E)   Plateau and baseline absorbance values have to be defined to correct for radial dilution [1].

(F)   Determination of position of point with greatest slope for time correction.

(G)   Leakage from any cell should be obvious.

(H)   A start- and endpoint for data to be included in the analysis has to be defined.

(I)   The editing routine has to be reproducible.

(K)   Possibility for exclusion of one or more scans from a series of scans in case of electronic noise etc. in a single scan.

---

[1] Radial dilution is an effect caused by the sector shape of the centrifuge cell. As solute molecules sediment to the bottom of the cell, where the cross-sectional area is larger, the solute molecules become more and more diluted. This effect can be observed when the absorbance of the plateau decreases for successive scans.

It is important to minimize any possibility for operator error. As more tasks are accomplished by the computer, the fewer errors can be introduced by the user. However, enough flexibility needs to be incorporated to allow for judicial choices in questionable situations. For example, a baseline might be distorted in a portion of its length because of a scratch in the cell window. The operator should then be able to exclude this region from baseline averaging. Or, the initial scan chosen for the determination of the point of steepest slope and baseline averaging should (1) include a long enough stretch of baseline to obtain a reasonable average and (2) should possess a sharp enough boundary to determine unambiguously the point of steepest slope. Since these parameters vary with each experiment, it is necessary to allow the user to make his own selection.

The editing algorithm was designed to integrate every requirement listed above. Only requirements (D), (E) and (F) were left for the user to be determined. Requirement (E) was assisted by plotting the derivative of the absorption data [*2] and (F) was satisfied by a prompt for visual inspection of the edited data.

2.1.3.2 van Holde - Weischet sedimentation velocity analysis

The van Holde and Weischet analysis method was implemented by first time correcting each scan. Time correction is necessary since during the initial period of acceleration of the rotor it is impossible to define a uniform centrifugal field strength. Since measurements are not taken during this period, time correction can be accomplished by observing a well defined point in the boundary (which will sediment with constant velocity once the rotor has obtained its final

---

[*2] The derivative of the absorption gradient cannot be used as a default to determine the steepest point in the slope, because in cases of large diffusion, the numerical derivative can be obliterated by the noise in the data and will not any useful results. In such a case the user has to determine the point of steepest slope unassisted. If the boundary is non-symmetric and sharp, the derivative will provide an excellent pointer to the location of the point of steepest slope.

speed) and extrapolating back to the time when it would have been at the meniscus.

Time correction in the van Holde - Weischet method is accomplished by taking the fractional distance {[absorbance(plateau) - absorbance(point of greatest slope in boundary)]/[absorbance(plateau) - absorbance(baseline)]} and finding the radial position in each scan corresponding to this fractional distance. The point of greatest slope (not necessarily the midpoint) will continue to sediment along this fractional distance and is chosen because the least amount of radial variation can occur around this point. The fractional distance has to be calculated for each scan anew because radial dilution will change the absolute height of this point from scan to scan. Based on this point, a time correction is calculated for each cell by linear least squares fitting and results from all cells are averaged, if there are more than one cell in the scan. A possibility to override the automatic time correction is also included, which allows the user to enter a hand-calculated time correction.

Next, the distance [absorbance(plateau) - absorbance(baseline)] is divided into *n* equal horizontal divisions along the included data, and the apparent sedimentation coefficients for each intersection with the data is calculated. If the number of divisions is too great, the shape of the boundary is approximated by extrapolating between adjacent datapoints. If more than one division falls on the extrapolated range, the analysis is aborted and a smaller number of divisions has to be used. Apparent sedimentation coefficients from each division are plotted for each scan against the inverse square root of the time of the scan. Apparent sedimentation coefficients corresponding to the same division number are fitted by a linear least squares fitting routine. The calculated line is extrapolated to infinite time to determine the uncorrected sedimentation coefficient for each species in the boundary.
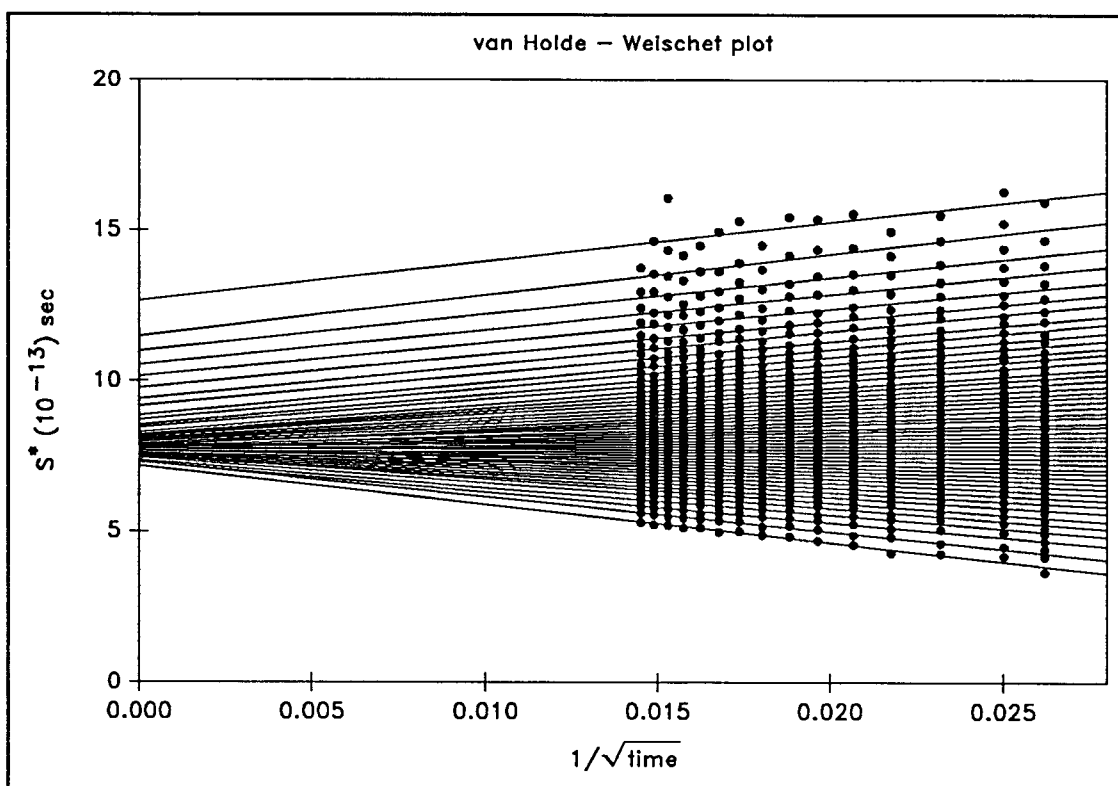
Figure 2.1.5: van Holde - Weischet plot for data shown in Figure 2.1.4.

If several lines converge in one or more points at infinite time, individual groups can be defined by grouping a range of intercepts and averaging their extrapolated S-values. Later, temperature and buffer corrected sedimentation coefficients for each group are calculated by standardizing sedimentation coefficients to conditions in water at 20°C [12]:

$$(9) \qquad S_{20,W} = \frac{(1 - \bar{V}\rho)_{20,W} \; \eta_{T,b}}{(1 - \bar{V}\rho)_{T,b} \; \eta_{20,W}} \; S_{T,b}$$

The results for the temperature and buffer corrected sedimentation coefficients for each group are displayed in Table 2.1.1:

```
Cell #1: CYT. C Oxidase (Triton X)

V-BAR for your Sample:                           .740
Density of your Buffer at 19.83 °C:             1.004
Viscosity of your Buffer at 19.83 °C:           1.143
Rotor Speed:                               38750.000

Calculated sedimentation coefficient from 1. group: 9.89901
Sedimentation coefficient corrected for water at 20°C from 1. group: 11.4274

Calculated sedimentation coefficient from 2. group: 7.71552
Sedimentation coefficient corrected for water at 20°C from 2. group: 8.90678
Second Moment Analysis for Run ID: 232, Cell #1
```

Table 2.1.1: Calculated results from van Holde - Weischet analysis of Figure 2.1.5. Two groups were chosen in this example by averaging the heterogeneous and homogeneous portion of the data shown above.

Also, a sedimentation distribution showing the fractional sedimentation coefficients for each extrapolated line is plotted. These distributions can be displayed both uncorrected and corrected. These graphs are obtained by plotting the intersects at $t = \infty$ of the van Holde - Weischet plot versus an arbitrary scale from 0% to 100%.

2.1.3.3 Second moment sedimentation velocity analysis

For the second moment method, time correction is implemented in the same way as for the van Holde - Weischet method. Second moments $\| R_s \|^2$ for each scan are calculated by the following method [13]:

$$(10) \qquad \| \overline{R_s} \|^2 = \frac{\sum_{i=R_m}^{R_p} R_i^2 * [A(R_i) - A(R_{i-1})] * [R_i - R_{i-1}]}{\sum_{i=R_m}^{R_p} [A(R_i) - A(R_{i-1})] * [R_i - R_{i-1}]}$$

where R_i is the radial position at point i, R_m is the radial position at the meniscus, and R_p the radial position of a user defined integration limit on the plateau. Second moment points and their distance from the point with greatest slope are listed. Sedimentation coefficients are calculated from each second moment point and listed as uncorrected and temperature, density
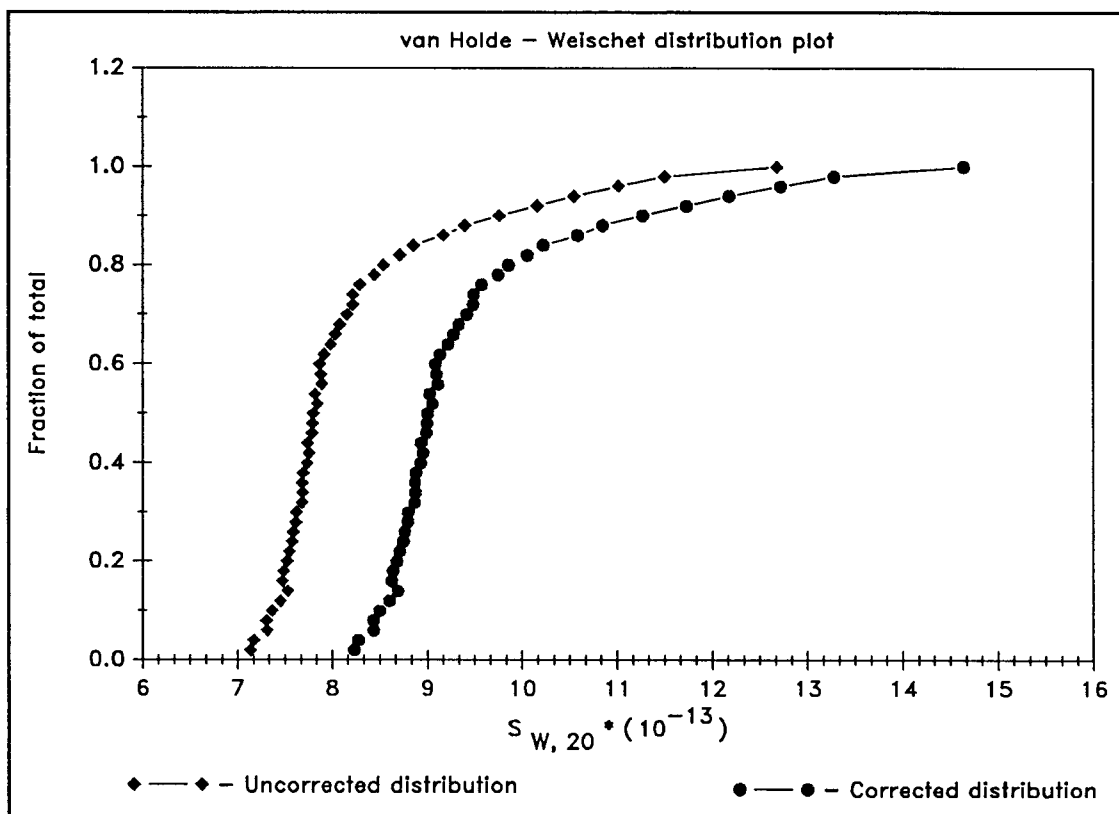
Figure 2.1.6: Sedimentation velocity data for cytochrome C oxidase, distribution plot. (Data courtesy of Dr. Jeff Hansen, University of Texas, San Antonio).

and specific volume corrected coefficients. Corrected and uncorrected sedimentation coefficients

from all scans of each cell are averaged and reported:

```
Cell #1: CYT. C Oxidase (Triton X) I=0.45


Scan:  Estm. Mpoint:  Calc. Mpoint:  Difference:  Calc. S-Val.:  Corr. S-Val.:

  1       6.106 cm       6.147 cm      - .041 cm      10.303         11.894
  2       6.126 cm       6.164 cm      - .037 cm      10.402         12.008
  3       6.146 cm       6.181 cm      - .034 cm       9.841         11.361
  4       6.164 cm       6.200 cm      - .036 cm       9.566         11.043
  5       6.184 cm       6.230 cm      - .047 cm       9.892         11.419
  6       6.202 cm       6.252 cm      - .050 cm       9.762         11.270
  7       6.221 cm       6.273 cm      - .052 cm       9.688         11.183
  8       6.240 cm       6.290 cm      - .050 cm       9.425         10.881
  9       6.260 cm       6.321 cm      - .061 cm       9.653         11.143
 10       6.277 cm       6.337 cm      - .060 cm       9.435         10.892
 11       6.296 cm       6.356 cm      - .060 cm       9.302         10.738
 12       6.316 cm       6.374 cm      - .057 cm       9.167         10.582
 13       6.336 cm       6.415 cm      - .079 cm       9.580         11.059
 14       6.354 cm       6.424 cm      - .071 cm       9.262         10.692
 15       6.370 cm       6.434 cm      - .064 cm       8.992         10.381

The average values for Cell #1 are:       - .053 cm       9.618         11.103
```

Table 2.1.2: Calculated results for the second moment analysis of data shown in Figure 2.1.4.

2.1.3.4 Equilibrium analysis

Data from sedimentation equilibrium experiments can be analyzed by a nonlinear least squares fitting method. The Levenberg-Marquardt method [9],[14] (program LEV-MAR.FOR, see Appendix C) is a robust method which optimizes a set of parameters by following the steepest gradient descent of the parameter's χ^2 surface. Initial guesses for each parameter are required as a starting point for the χ^2 minimization. The Jacobian matrix of the parameter space is computed and each parameter is adjusted in the negative direction of the gradient.

Initial guesses for the parameters are obtained as follows: The baseline is estimated by averaging a number of datapoints between the reference and sample meniscus. After subtracting the baseline voltage from all datapoints, all other parameters are calculated by a linear least squares fit of all datapoints to the following function:

$$(11) \quad \ln(V_i) - \ln(V_a) = \frac{\omega^2(1 - \bar{V}\rho)\, M_{wr}}{2RT} * (r_i^2 - a^2)$$

where V_i is the voltage at point i, V_a is the voltage at the meniscus, r_i is the radial position of point i, a is the position of the meniscus, ω is the radial velocity, M_{wr} is the weight-average molecular weight of the mixture at point r, $\bar{v}$ the specific volume, ρ the density of the solvent, R the gas constant and T the absolute temperature. A χ^2 value is calculated and minimized by varying the baseline about the initial guess (program BLINE.FOR [15], see Appendix B), each time subtracting the baseline voltage from each datapoint. An optimal baseline is obtained when the χ^2 is minimal. The optimized guesses are used for initial starting parameters in the Levenberg-Marquardt nonlinear fitting routine. For a single component fit, the ln(A) versus r^2 plot can be graphed. Although absorbance values are not plotted, voltages are directly proportional to the absorbance and can be calibrated using the stairsteps method outlined the "Experimental setup" section (2.1.1.1).

Figure 2.1.5 shows the Ln(V) versus r^2 plot for a uracil glycosylase inhibitor complex (data kindly provided by Dr. Ken van Holde and Sam Bennett, Oregon State University). Conditions of the runs were as follows: Temperature: 19.5°C, rotor speed: 20,000 rpm, wavelength: 280 nm, buffer: 30 mM Tris-HCl pH 8.0, 1 mM EDTA, 1 mM DTT, 100 mM NaCl, 5% glycerol. The density corrections for this buffer is 1.029.

The quality of a fit is assessed by plotting the residuals of the fit. An example is shown for the single component fit in Figure 2.1.6 for the data shown in Figure 2.1.5. If the residuals are distributed about the zero line in a random fashion, the fit is considered to be good and indicative of a single component system. If the residuals follow an oscillating pattern about the fit, a multicomponent system is likely the cause. If the value for the baseline is too low, the residuals follow a trace which intersects the zero line twice.

For equilibrium experiments it is possible to average up to 20 scans of the same experiment in one analysis. For each scan an individual molecular weight as well as an average of all previous scans is displayed in a file (Table 2.1.3).
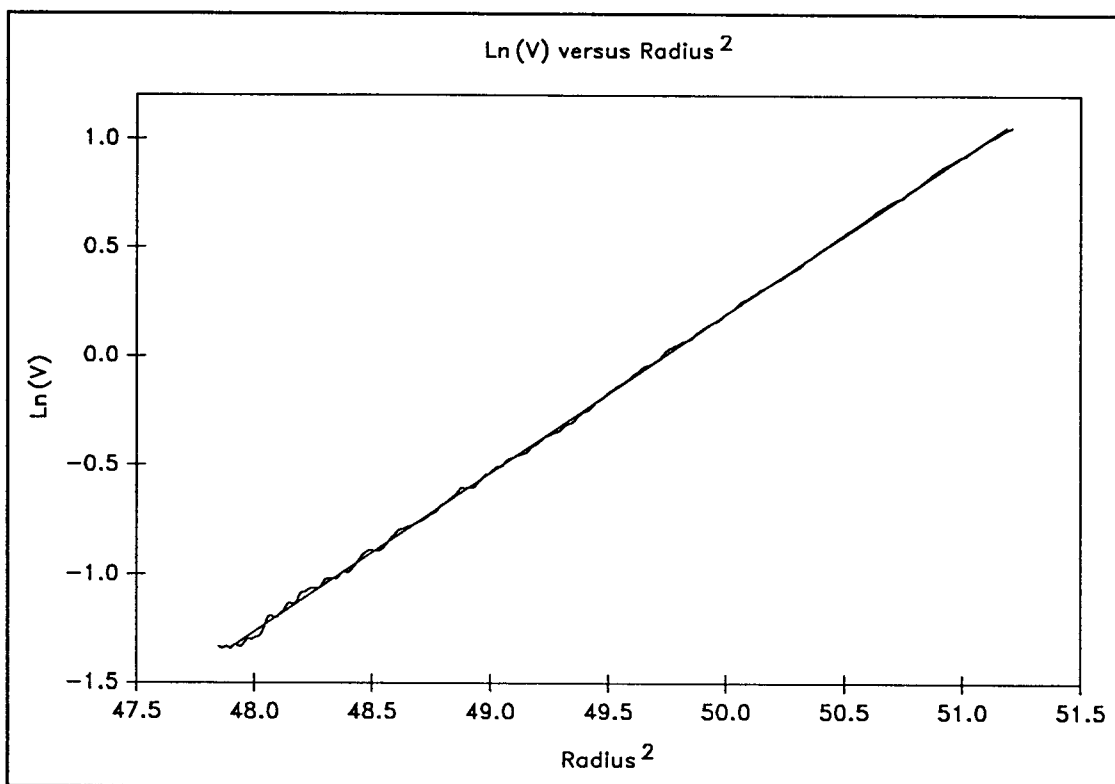
Figure 2.1.7: Ln(V) versus r^2 plot for sedimentation equilibrium data of uracil-glycosylase inhibitor complex. (Data courtesy of Dr. Ken van Holde and Sam Bennett, Oregon State University).
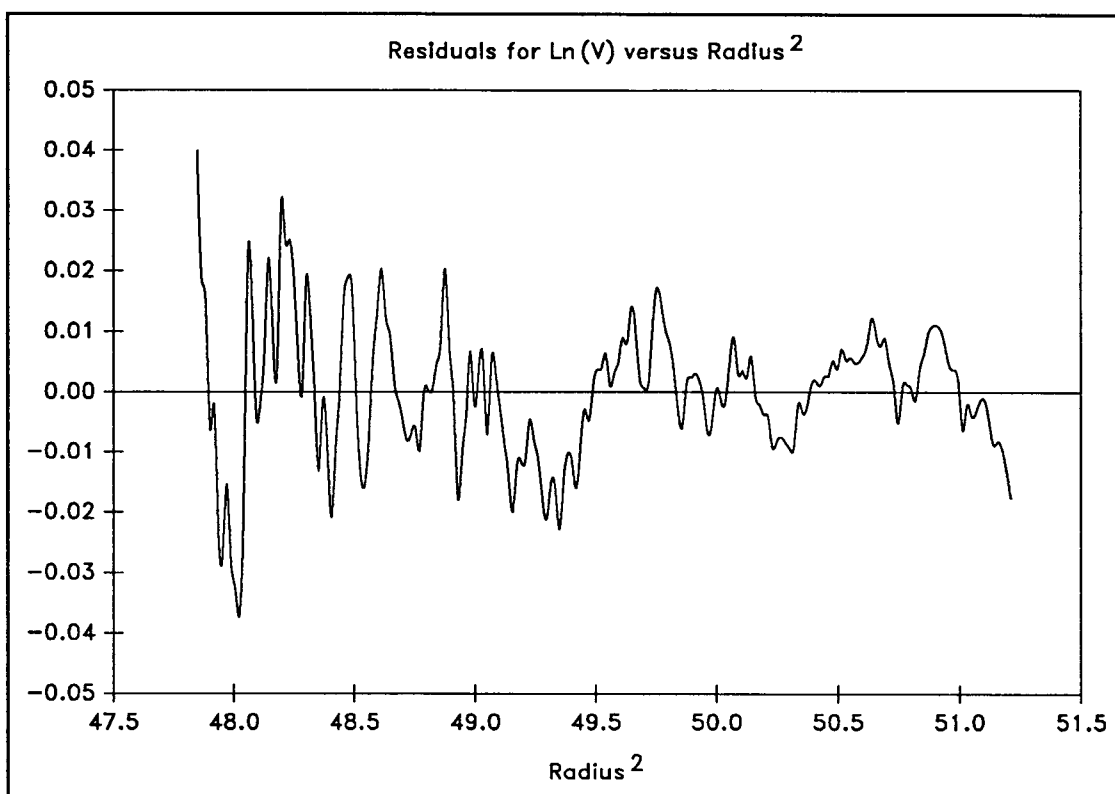
Figure 2.1.8: Residuals for single component fit on data shown in Figure 2.1.6.

```
*****************************************************************
  EQUILIBRIUM ANALYSIS FOR RUN 5186b, CELL #1 :
*****************************************************************

  Cell #1: klh4-1

  V-BAR of Sample:                        .740
  Density of Buffer at 14.74 °C:         1.029
  Viscosity of Buffer at 14.74 °C:       1.146
  Rotor Speed:                       20,000.000


  Statistics for Scan # 1 :        Molecular Weight:

  Slope:            .7177          Corr.   MW: 3.2818e+04 g/mol
  Intercept: -   35.6912          Average MW: 3.2818e+04 g/mol
  Correlation:      .9999          Uncorr. MW: 7.8283e+03 g/mol
  Chi_Square:       .0280


  Statistics for Scan # 2 :        Molecular Weight:

  Slope:            .7254          Corr.   MW: 3.3170e+04 g/mol
  Intercept: -   36.0871          Average MW: 3.2994e+04 g/mol
  Correlation:      .9998          Uncorr. MW: 7.9124e+03 g/mol
  Chi_Square:       .0323


  Statistics for Scan # 3 :        Molecular Weight:

  Slope:            .6997          Corr.   MW: 3.1994e+04 g/mol
  Intercept: -   34.7598          Average MW: 3.2660e+04 g/mol
  Correlation:      .9998          Uncorr. MW: 7.6318e+03 g/mol
  Chi_Square:       .0379


  Statistics for Scan # 4 :        Molecular Weight:

  Slope:            .7273          Corr.   MW: 3.3256e+04 g/mol
  Intercept: -   36.1721          Average MW: 3.2809e+04 g/mol
  Correlation:      .9999          Uncorr. MW: 7.9329e+03 g/mol
  Chi_Square:       .0249
```

Table 2.1.3: Calculated results for equilibrium run of data in Figure 2.1.5. Results for four scans are shown, the underlined number is the final molecular weight average of all measured scans.

2.2 Quasi-elastic light scattering

The determination of diffusion coefficients by quasi-elastic light scattering is a convenient method to complement data obtained from sedimentation experiments. Sedimentation and diffusion coefficients are related by following well-known formula:

$$(12) \qquad \frac{S_0}{D_0} = \frac{M*(1 - \bar{v}*\rho)}{RT}$$

where M is the molecular weight, $\bar{v}$ the specific volume, ρ the density of the solvent, R the gas constant and T the absolute temperature. Therefore, having obtained the diffusion coefficient D_0 and the sedimentation coefficient S_0, it is possible to calculate the molecular weight (given

the specific volume) or the specific volume (given the molecular weight).

2.2.1 Experimental setup

Laser light from a 488 nm argon laser is passed through a temperature controlled sample (within 0.1°C) and the intensity of scattered light is measured at a 90° angle from the laser beam. The fluctuation of intensity of the scattered light is proportional to the diffusion of molecules through a small volume element of solvent. Molecules passing through this volume element scatter light which can be measured with a photomultiplier tube. Photons are counted and an autocorrelation function is computed by a hardware implemented autocorrelator.

The autocorrelation functions is a measure of the rapidity of fluctuation of the scattered light intensity. Time increments of varying size from the sample time to a time slice 127 times the length of the sample time are checked for autocorrelation by calculating following average, called the autocorrelation function:

(13)
$$A(\tau) = \overline{\Delta i(t) \Delta i(t + \tau)}$$

where $\Delta i(t)$ is the difference in intensity from the average intensity [16],[17].

The correlator is driven by a PC computer where the data is analyzed. After sufficient counts are obtained, the autocorrelation function can be fitted to following function:

(14)
$$A(a_i, b_i, c, \tau) = \sum_{i=1}^{k} [a_i * e^{b_i * \tau}] + c$$

(15)
$$with: \quad b_i = D_i * \frac{8 \pi^2 n^2}{\lambda_0^2} * (\sin \frac{\theta}{2})^2$$

where a_i, b_i and C are adjustable parameters and a_i corresponds to the amount of component i, b_i is proportional to the frequency of component i, C is the baseline, k the
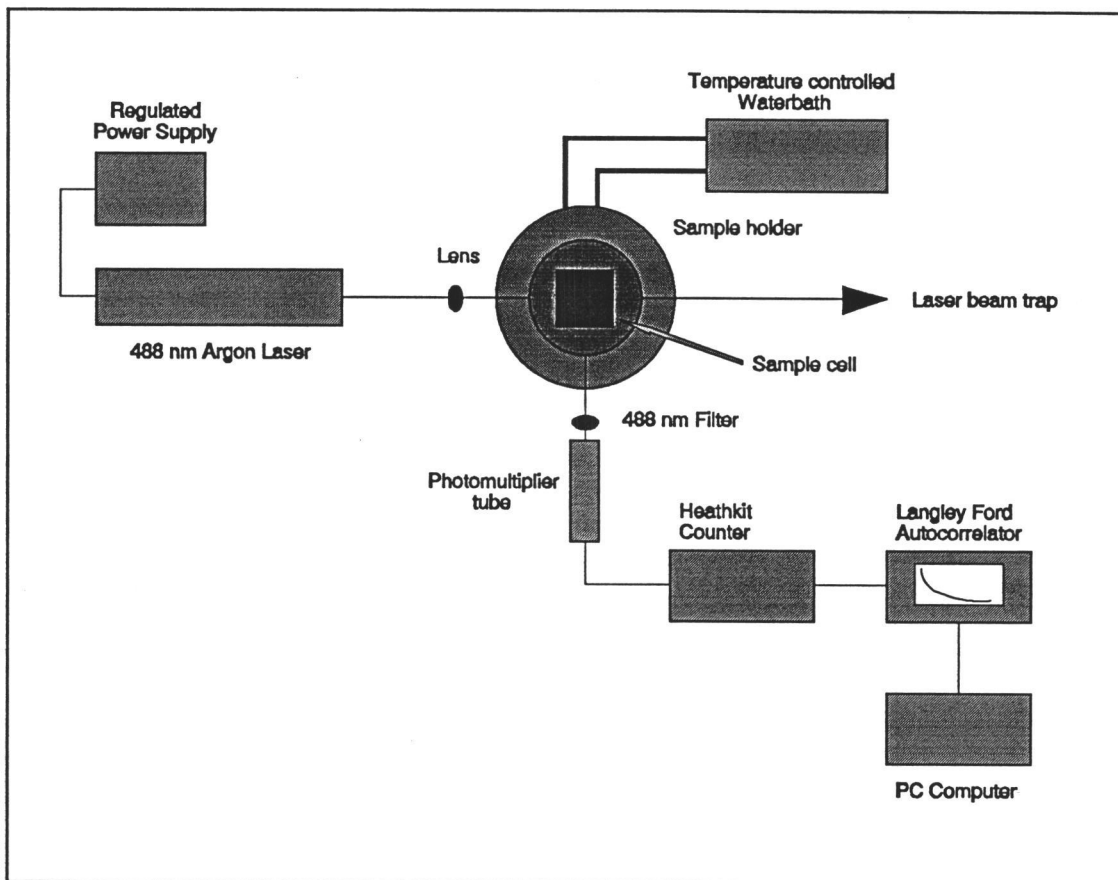
Figure 2.2.1: Experimental setup for quasi-elastic light scattering experiments for diffusion studies.

maximum number of components to be fitted, τ the sample time and A the correlation at sample time τ. The diffusion coefficient D_i for component i is calculated from b_i, where n is the refractive index, λ_0 the wavelength of the laser, and θ is the scattering angle.

Although a single component fit can be obtained by hand through linear least squares fitting (a baseline estimate is provided by the autocorrelator), this method is inaccurate and laborious. Since a nonlinear fitting procedure cannot be reasonably calculated by hand, computer implementation is again essential.

A program (LaserScan) was developed to (1) drive the Langley-Ford autocorrelator, (2) to acquire the autocorrelation data and (3) to analyze the data by the Levenberg-Marquardt nonlinear fitting routine for single and double component datafits. The program consists of

several independent modules, which are driven by the main program, LASER.EXE. A complete

description of the program is included in Appendix C. The program is designed to automate as

much as possible the process of data acquisition and data analysis. Following is an overview

of program structure and module arrangement:
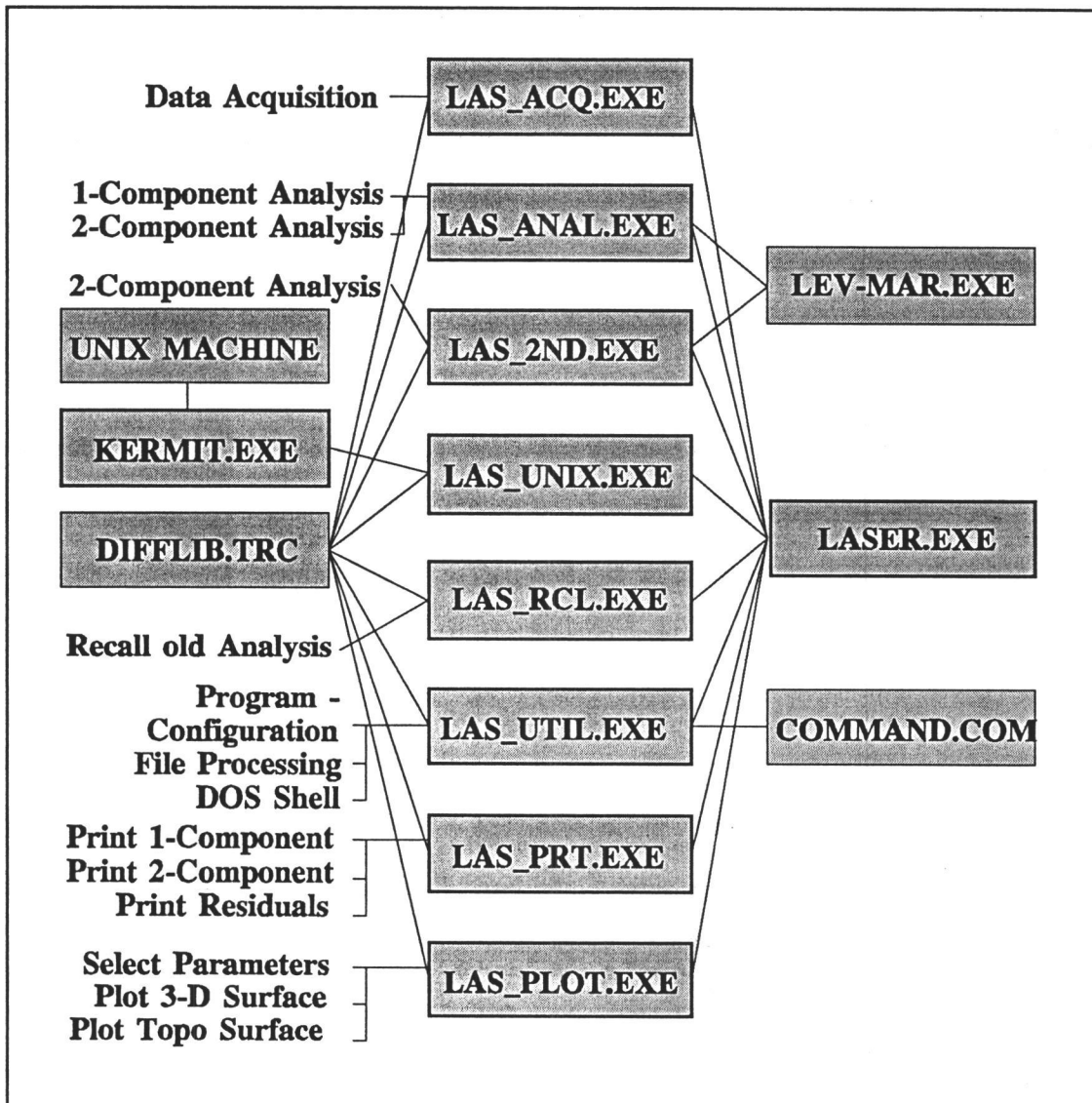
Figure 2.2.2: Program structure of LaserScan. Shown are individual modules (thick borders) and their dependencies (thin borders). Lines represent pathways of accessing the subprograms. The program is entered through LASER.EXE.

2.2.2 Data acquisition

Data acquisition to the Langley-Ford autocorrelator was accomplished through serial connection with the PC computer. The Langley-Ford autocorrelator provides a RS-232 interface capable of bidirectional communication. This allows for a driver program which can set all parameters from the computer, rather than manually on the keypad of the autocorrelator. The driver program was written to allow a cycling mode which cycles stepwise through a series of predetermined sample times, until a user selectable threshold level of counts for channel 1 has been obtained. The data is written to a file together with the parameters for the run. Each scan contains the following information:

1. Temperature of sample during acquisition

2. Sample time of run

3. A flag indicating if the last 16 channels were delayed for baseline approximation

4. Total time elapsed during data acquisition

5. Total sample time counts

6. Total shift register counts

7. Total add commands

8. Number of overflow counts (dust particles and air bubbles in laser beam produce overflow counts. Also, if the intensity for the laser beam is too high, overflow counts are registered. Overflow counts add to the background and should be avoided.

9. Counts for 128 (or 144, without baseline) sample time channels.

10. 16 delayed channels for baseline approximation.

The user can determine if a new temperature should be entered for each sample time, in case the temperature control is not stable. During data acquisition, the driver program will reset automatically all parameters on the autocorrelator for each new sample time. All registers will be reset to 0, the sample time is set to the new value, and a new sample time cycle is acquired.

2.2.3 Data analysis

To fit the datapoints to the autocorrelation function, the nonlinear Levenberg-Marquardt χ^2 minimization algorithm is employed, because the baseline term prevents linearization. Although a baseline estimate is provided by the autocorrelator (the last 16 channels are substantially delayed), the estimate is often not accurate enough.

Since the Levenberg-Marquardt method requires initial guesses for the gradient descent, starting points for each adjustable parameter have to be estimated. The last sixteen channels are delayed and their average is a convenient initial estimate for the baseline. This estimate is subtracted from each datapoint. If the difference becomes negative for any point in the data (because of improper baseline estimation), the value of the baseline average is incrementally decreased until the difference is positive for each point. This procedure assures that during linearization the logarithm for each datapoint is defined. Once an appropriate baseline has been found, the differences are fitted to a linearized version of the autocorrelation function by linear least squares:

(16) $$\ln\left[A(a, b, \tau) - c\right] = \ln(a) + b*\tau$$

Parameters a and b serve as initial guesses for parameters a_i and b_i in equation 40 with $n = 1$ for a single component, nonlinear fit (nonlinear, because the baseline will be readjusted as well). After calculation of all parameters for the single component fit, the residuals for the single component fit are displayed. If an oscillating pattern of the residuals about the fit exists, a double component fit can be calculated. Again, initial guesses for all parameters have to be estimated.

For the double component fit, $n = 2$ and initial estimates for a_1, a_2, b_1 and b_2 have to be obtained. By visually inspecting the trace of residuals for the single component fit, the user can split the spectrum in 2 not necessarily equally sized sections. Each section is individually linearized and fitted by a linear least squares procedure, after the modified baseline estimate

is subtracted from each datapoint. Parameters obtained from the linearized fit can then be used as initial guesses for the Levenberg-Marquardt double component fit. Since the double component fit allows for 2 additional degrees of freedom, a lower χ^2 value can almost always be obtained.

A routine is included (LAS_2ND.EXE) which allows analyzing the double component fit exclusively, after a single component has been successfully fitted to the dataset. This routine allows the user to enter a different starting point for the 2-component fit by dividing the spectrum into 2 differently sized initial sections each time the routine is run. This will give a different starting point for parameters a_i and can result in a modified set of parameters due to the various starting positions in the parameter space for the gradient descent. Sometimes, the routine has to be run several times until a sufficiently low value for the χ^2 has been obtained.

Once a dataset of several sample times has been analyzed, the results are written to a file:

```
*************************************************************************
  RESULTS FOR FILE: 208_24, pPOL-1 208-24 plasmid DNA, Hind III digest
*************************************************************************

  Sample time:          5.0000e-06 sec
  Temperature:          3.60 °C
  Wavelength:           488.00 nm
  Viscosity:            1.5869e+00 cp
  Scattering angle:     90.00 degrees
  Refractive index:     1.330
  Total time of run:    1.1001e+03 sec
  Sample time count:    2.1992e+08
  Overflow count:       8.6800e+02
  Shift reg. pulses:    5.0419e+08

  Laserscan Results and Statistics:
  ─────────────────────────────────────────────────────────────────────
  Data for single component fit:   A(τ)=K1*exp(K2*τ)+K3
  ─────────────────────────────────────────────────────────────────────

  Diffusion Coeff.:     3.3798e-08 cm²/sec
  Corrected Diff.:      5.6701e-08 cm²/sec

  Amplitude (K1):       8.05517e+05   Stand. dev. σ(K1):   8.12458e-04
  Frequency (K2):      -3.96438e+03   Stand. dev. σ(K2):   4.66350e-06
  Baseline  (K3):       9.91526e+06   Stand. dev. σ(K3):   5.02620e-04
  Estimated Baseline:   9.80086e+06

  Statistical Tests:

  F — Test:             0.00000e+00   Total Iterations:       1388
  Chi — Square:         4.25263e+07   Calculation Time:        .00 sec
  ─────────────────────────────────────────────────────────────────────
  Data for double component fit:   A(τ)=K1*exp(K2*τ)+K3*exp(K4*τ)+K5
  ─────────────────────────────────────────────────────────────────────

  Diffusion Coefficient for 1. Component:        1.3078e-07 cm²/sec
  Diffusion Coefficient, corrected to (20°,W):   2.1940e-07 cm²/sec
  Relative Amount of 1. Component:               37.15 %

  Diffusion Coefficient for 2. Component:        2.6187e-08 cm²/sec
  Diffusion Coefficient, corrected to (20°,W):   4.3932e-08 cm²/sec
  Relative Amount of 2. Component:               62.85 %

  Amplitude-1 (K1):     3.79874e+05   Stand. dev. σ(K1):   1.71803e-03
  Frequency-1 (K2):    -7.67004e+03   Stand. dev. σ(K2):   4.25448e-05
  Amplitude-2 (K3):     6.42644e+05   Stand. dev. σ(K3):   5.49662e-04
  Frequency-2 (K4):    -1.53580e+03   Stand. dev. σ(K4):   2.35558e-06
  Baseline (K5):        9.72196e+06   Stand. dev. σ(K5):   8.50730e-04

  Statistical Tests:

  F — Test:             0.00000e+00   Total Iterations:       1621
  Chi — Square:         1.35800e+07   Calculation Time:     282.70 sec
```

Table 2.2.1: Results for a multicomponent fit of light scattering data for a single sample time.

Also, results from all sample times for one particular run are summarized in a single file:

```
       Sample: 208_24, pPOL-1 208-24 plasmid DNA, Hind III digested

*********************************************************************************
 Corrected diffusion coefficients for sample time: .000008 sec

 Single Component fit:        9.85803e-08 cm²/sec (100.00%) (X² = 6.00854e+07)
 Double Component fit: (1)    1.53908e-07 cm²/sec ( 55.08%) (X² = 1.43018e+07)
                       (2)    3.68427e-08 cm²/sec ( 44.92%)
 Weighted Average: (1 & 2)    1.01322e-07 cm²/sec
*********************************************************************************
 Corrected diffusion coefficients for sample time: .000009 sec

 Single Component fit:        9.52576e-08 cm²/sec (100.00%) (X² = 6.93228e+07)
 Double Component fit: (1)    1.51106e-07 cm²/sec ( 56.39%) (X² = 1.48697e+07)
                       (2)    3.72769e-08 cm²/sec ( 43.61%)
 Weighted Average: (1 & 2)    1.01469e-07 cm²/sec
*********************************************************************************
 Corrected diffusion coefficients for sample time: .00001 sec

 Single Component fit:        9.26219e-08 cm²/sec (100.00%) (X² = 1.07075e+08)
 Double Component fit: (1)    1.43590e-07 cm²/sec ( 59.62%) (X² = 1.49713e+07)
                       (2)    3.58240e-08 cm²/sec ( 40.38%)
 Weighted Average: (1 & 2)    1.00070e-07 cm²/sec
*********************************************************************************
 Corrected diffusion coefficients for sample time: .000011 sec

 Single Component fit:        9.07247e-08 cm²/sec (100.00%) (X² = 1.26088e+08)
 Double Component fit: (1)    1.39980e-07 cm²/sec ( 61.77%) (X² = 1.75808e+07)
                       (2)    3.59498e-08 cm²/sec ( 38.23%)
 Weighted Average: (1 & 2)    1.00213e-07 cm²/sec
*********************************************************************************
 Corrected diffusion coefficients for sample time: .000012 sec

 Single Component fit:        8.84474e-08 cm²/sec (100.00%) (X² = 1.30617e+08)
 Double Component fit: (1)    1.34285e-07 cm²/sec ( 63.84%) (X² = 1.76445e+07)
                       (2)    3.58649e-08 cm²/sec ( 36.16%)
 Weighted Average: (1 & 2)    9.86988e-08 cm²/sec
*********************************************************************************
 Corrected diffusion coefficients for sample time: .000013 sec

 Single Component fit:        8.65482e-08 cm²/sec (100.00%) (X² = 1.98539e+08)
 Double Component fit: (1)    1.33588e-07 cm²/sec ( 64.38%) (X² = 2.60116e+07)
                       (2)    3.62039e-08 cm²/sec ( 35.62%)
 Weighted Average: (1 & 2)    9.88978e-08 cm²/sec

 etc...
```

Table 2.2.2: Summary of results for all sample times of one experiment.

Relative amounts are calculated for each component fitted in the double component fit. When sample times are increased, the amount of correlation for slower components will increase, while correlation for faster diffusing components will decrease. The reason is, that when the sample time is increased, slower components are allowed to contribute to the spectrum more and more.

Faster components, which contribute more to the shorter sample time portion of the

spectrum, will decrease with their total contribution. This is demonstrated in the next graph, which shows the differences in relative amounts for each fitted component for various sample times:
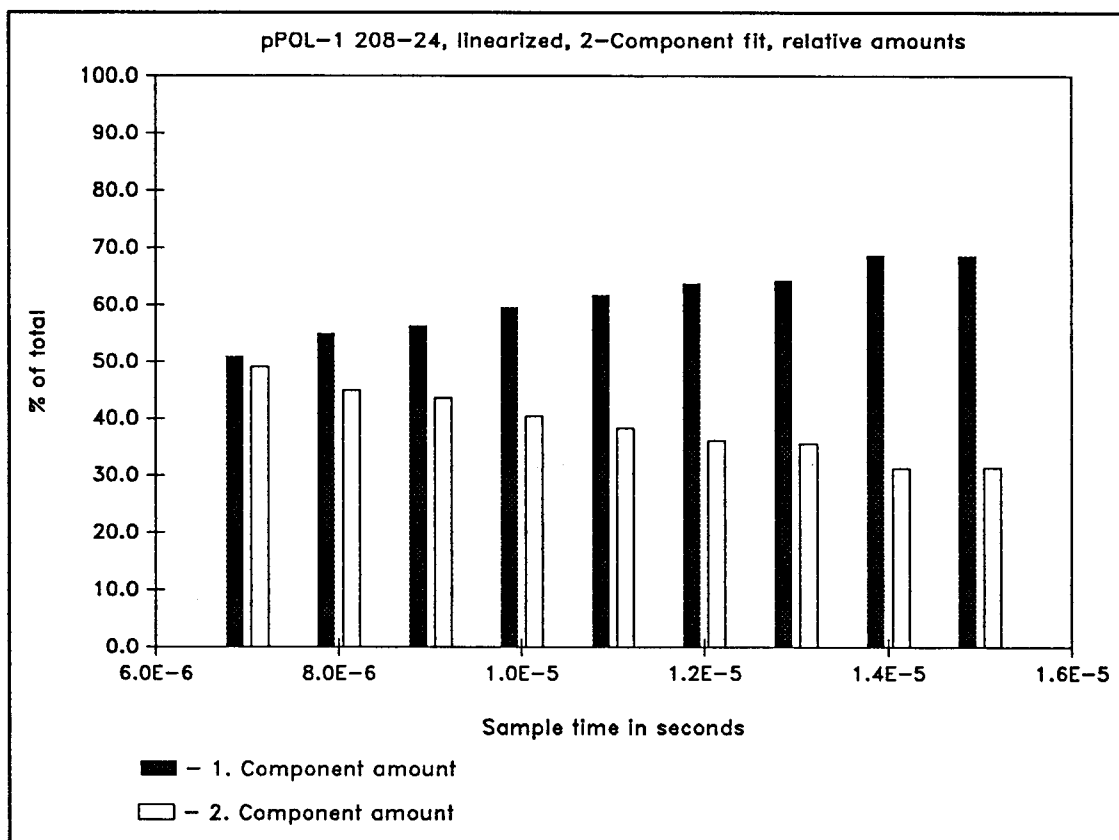
Figure 2.2.3: Relative contributions of both components to the total spectrum at various sample times. The relative amount of the faster component (component 2) decreases, while the slower component gains with increasing sample time.

Diffusion coefficients calculated from the fit are viscosity corrected for the temperature of the run. Four different diffusion coefficients are calculated:

1. Single component fit

2. Double component fit (1. component)

3. Double component fit (2. component)

4. Weighted average of both coefficients for the double component fit.

The weighted average of both double component coefficients is often quite similar to the single component coefficient, since the single component represents the best average for all components in the system. While the double component fit often provides a higher resolution of the data, a system rarely approaches the two ideal situations (single or double component). The double component fit provides a better resolution by virtue of 2 additional degrees of freedom. Almost always a lower χ^2 value can be obtained with the double component fit, although occasionally the χ^2 minimization might become stuck in a local minimum (discussed below).

If the system is not a true double component system, it is likely that one or both reported diffusion coefficients represent an average of a portion of the components in the system, just like the single component fit represents an weighted average of all components in the system. If one component is prevalent, it often is well isolated from the others, which are then averaged by the other component. This becomes apparent when diffusion coefficients for both components are plotted against the sample time. A well isolated coefficient will remain constant throughout the varying sample times, while an average of several components will change with sample time. Generally, the diffusion coefficient will decrease, as the sample time is increased, since more and more slower components (mainly dust and other large impurities) will contribute to the spectrum:
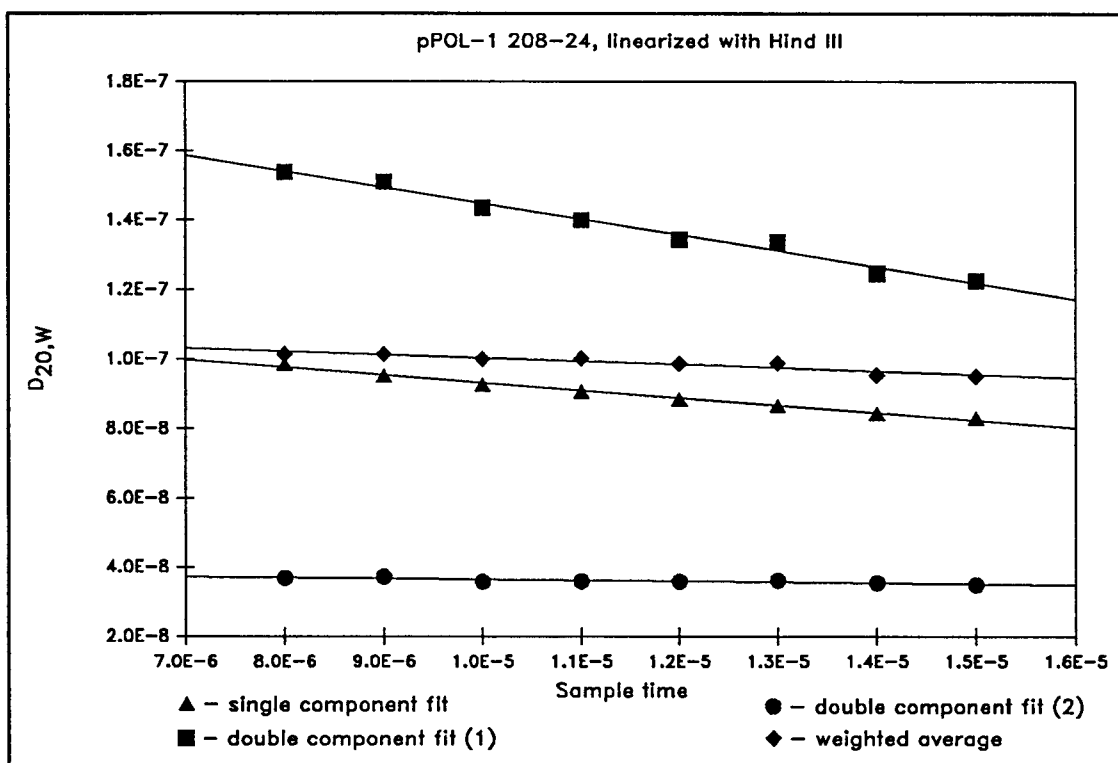
Figure 2.2.4: Diffusion coefficients plotted vs. increasing sample time. Coefficients are shown for the single component fit, both components of the double component fit and the weighted average of the double component fit.

2.2.4 Discussion of the Levenberg-Marquardt routine

While the Levenberg-Marquardt χ^2 minimization routine provides a robust optimization algorithm for nonlinear models, disadvantages exist. The same χ^2 value can be obtained by different combinations of parameters. A χ^2 three dimensional surface and a χ^2 topographic surface can be constructed by simultaneously varying two of the obtained parameters by several standard deviations while keeping all others constant, and calculating the resulting χ^2 value for a two dimensional grid. The resulting χ^2 surface can then be used to visualize the confidence limits for each parameter with respect to another (for a review, see [18]).

The shape of the valleys in the the topographic plot shown in Figure 2.2.5 visualizes the confidence in the parameter for which the plot has been calculated, a small change in χ^2

signifies a low confidence, a large change over the entire range with the valley in the center

signifies that the best possible parameters have been found. Such a situation is represented

in Figure 2.2.5:
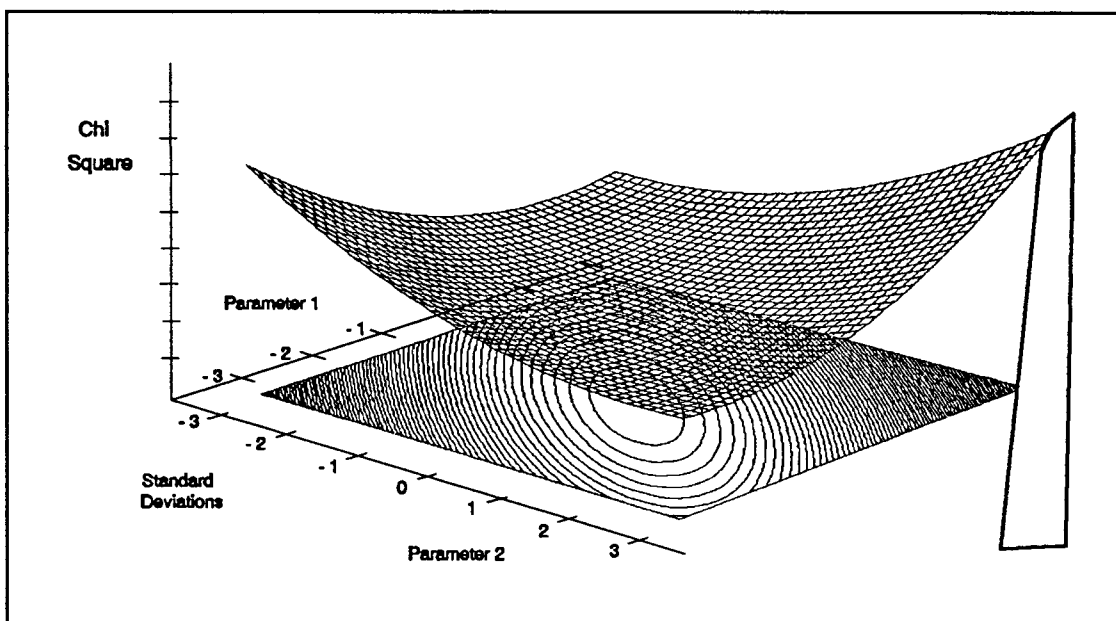
Figure 2.2.5: Topographic and three-dimensional map of parameters 1 and 2 for a single component nonlinear fit for pPOL-1 208-24 diffusion data by the Levenberg-Marquardt method. Parameters were varied by +/- 3 standard deviations.

Another problem associated with Levenberg-Marquardt optimization is the possibility of

a χ^2 surface with many local minima. If a local minimum exists, the routine may become stuck

in such a minimum, and the optimal solution is not approached. This will result in a fit which

obviously does not coincide with the data. In such a case, a new starting estimate might be able

to force the minimization on another path and help avoid the local minimum. This can be

accomplished by adding modifying the starting guesses by approximately 0.5% to 5.0% of their

absolute value. Initial guesses for parameters b_i can be varied more by selecting a different

breakpoint for the bisection of the spectrum.

A fit is considered to be good if the residuals are scattered about the zero line in a

random fashion. In the following examples, the residuals are plotted for the single and double component. The single component fit exhibits a sinusoidal pattern about the zero line, suggesting a multicomponent system which cannot be approximated by a single component model. However, the residuals for the double component fit exhibit a more random distribution of the residuals (and they are smaller because of a lower χ^2 value) about the zero line, suggesting that a double component model can approximate the real situation much better. While overlaying single and double component fits over the data does not reveal the differences in quality of fit (Figure 2.2.6), a plot of the residuals can highlight the difference in goodness of fit much better (Figure 2.2.7).
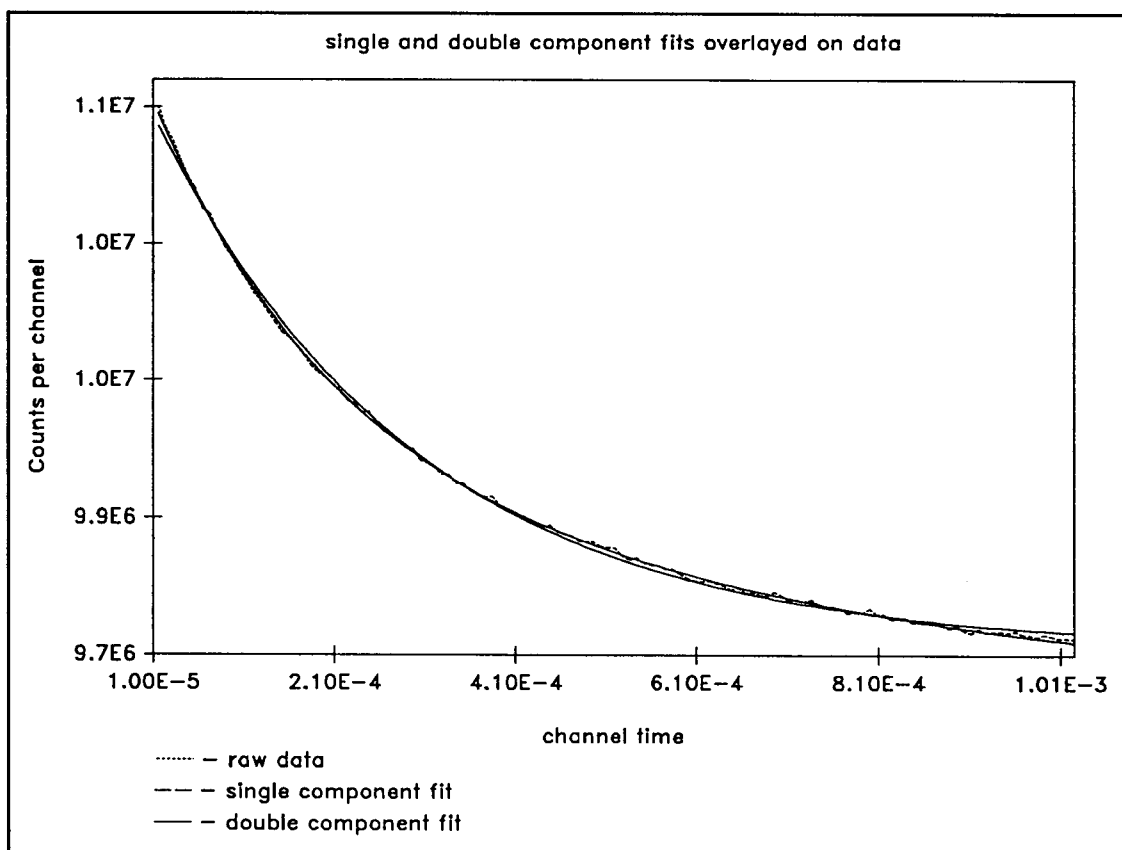
Figure 2.2.6: Data for pPOL-1 208-24 showing the raw data (sample time $8*10^{-6}$ seconds) overlayed on the single and double component fit.
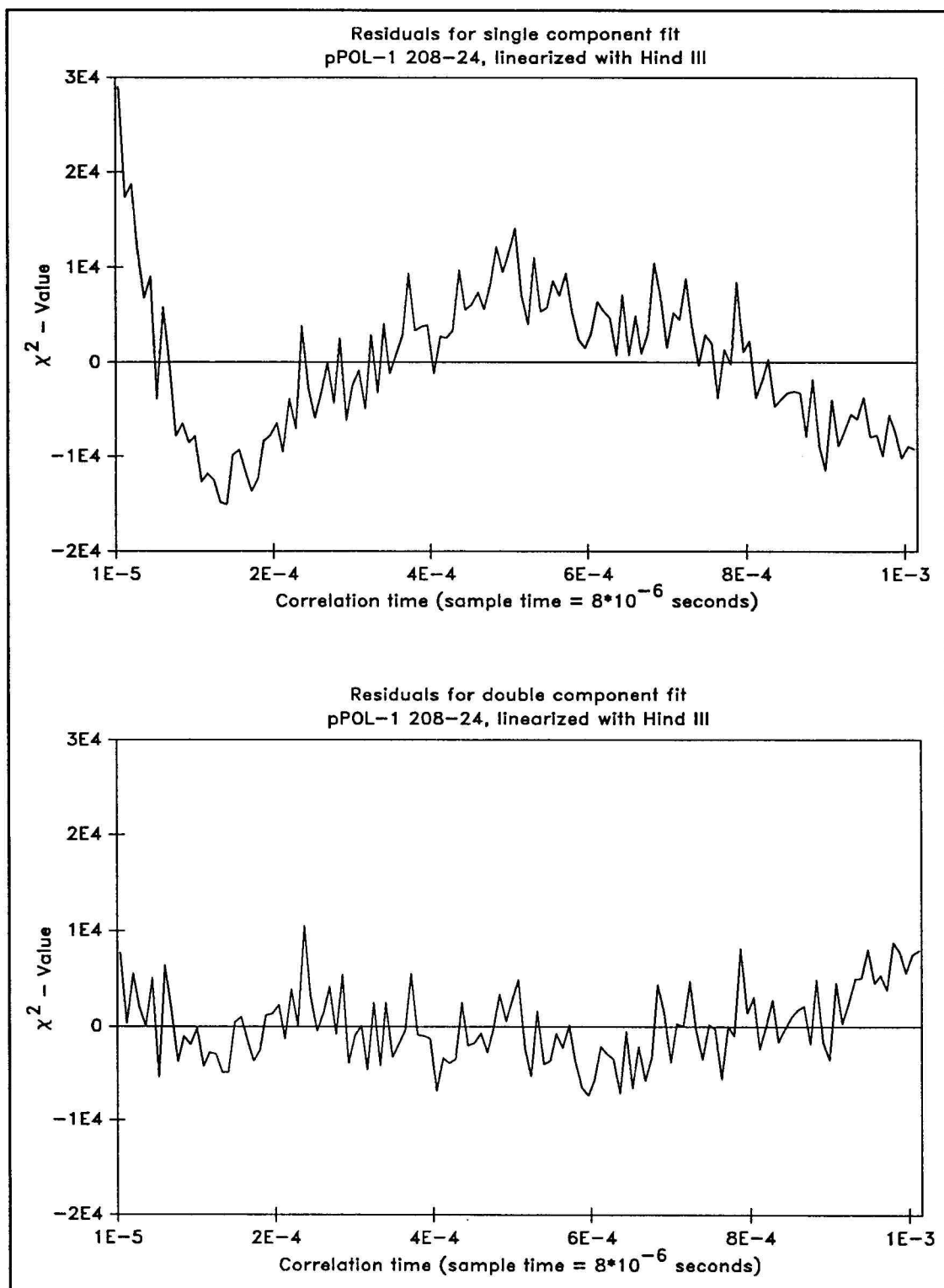
Figure 2.2.7: Residuals for single (above) and double (below) component fit. Notice that the double component fit has much smaller residuals. Data for pPOL-1 208-24, linearized.

Lastly, it should be mentioned that the quality of the fit to the dataset is not exclusively determined by the capability of the fitting routine, but also by the adequacy of the model the dataset is to be fitted to. While a sum of exponential terms plus a linear term for the baseline might approach an ideal system, nonideal interactions between components cannot be accounted for.

Furthermore, impurities, partially degraded materials, dust etc. might complicate the system to be fitted, and much more than just two components should be included in the model. For example, in order to appropriately fit the light scattering data for the pPOL-1 208-24 plasmid, a double component fit was required. The first component accounted for one or more components diffusing with a much larger diffusion coefficient, suggesting the presence of degraded material, such as nucleotides and shorter pieces of DNA. Another result supports this hypothesis: As sample times were increased, the relative amount of the second component increased. Since longer sample times emphasize the slower components present in the system, the first component (greater diffusion coefficient) was probably fitted to smaller material present in the sample.

The number of datapoints and the resolution of the data, however, is not sufficient to warrant a fit to more than two components, and higher component models are meaningless.

3. PLASMID CONSTRUCTION

3.1 Introduction

Current research in our laboratory is focussing on a number of aspects of chromatin structure and function. Efforts are underway to investigate transcription of reconstituted chromatin circular templates, the effects of reconstitution on topology and on the higher order structure of chromatin. The need for a versatile in vitro system for such studies is obvious. To address the many requirements for such a system a new plasmid vector system was created, called the pPOL-1 208-n series of plasmids. Some of the demands on this system are listed below:

1. Incorporation of a RNA polymerase I promoter site

2. Variability in size and topology

3. High efficiency of amplification to obtain large quantities for analytical experiments

4. Incorporation of a well-studied nucleosome positioning sequence, the sea urchin ribosomal 5S sequence

5. Minimal amount of extraneous vector DNA

6. Restriction site selection should allow for:

    a) separation of plasmid DNA from insert DNA with one enzyme (without cutting anywhere else)

    b) separation of promoter from repeat without cutting anywhere else

    c) use of low-cost endonucleases

A flowchart for the design of this system is shown in Figure 3.1.
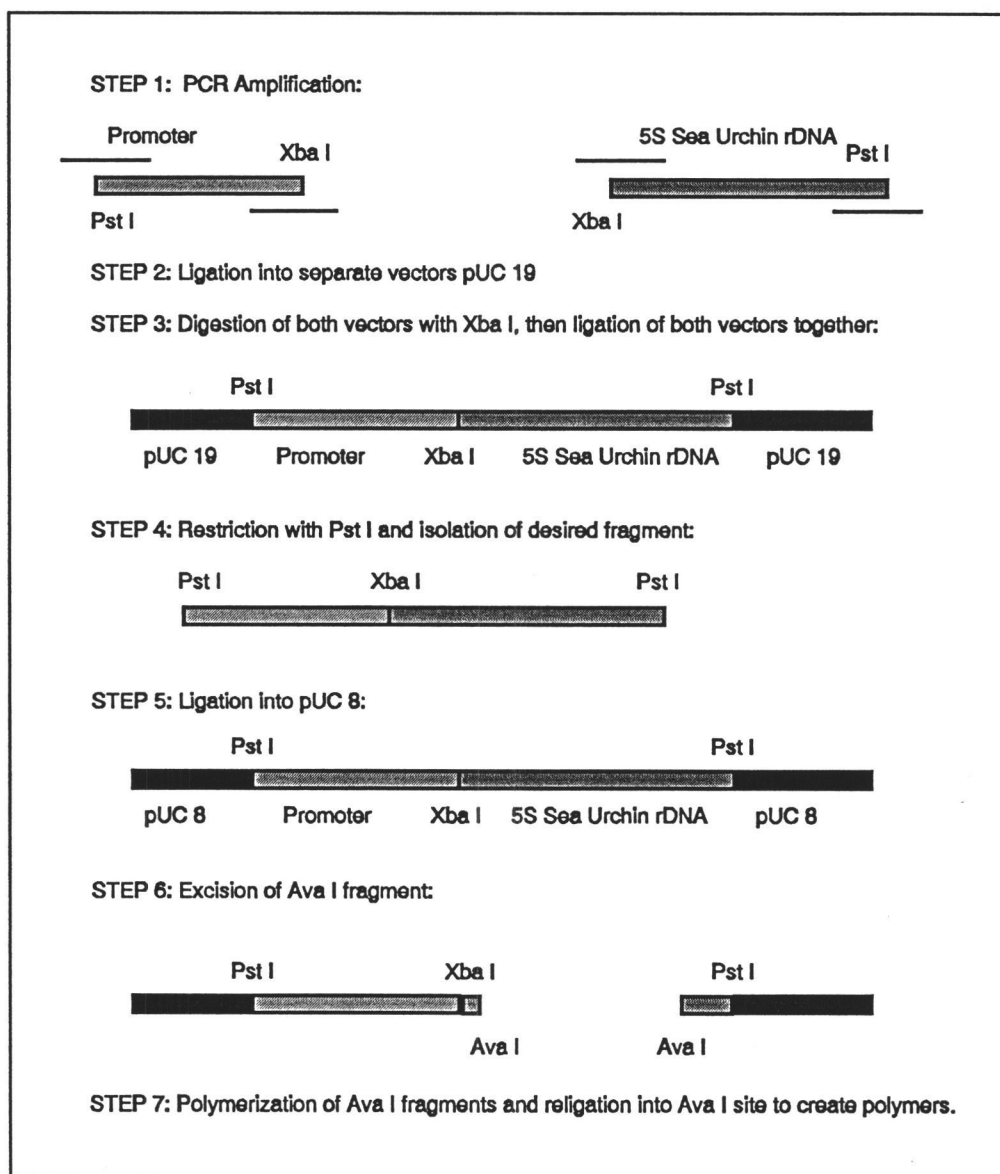
Figure 3.1: Flowchart for construction of pPOL-1 208-n plasmid series.

3.2 Monomer construction

The plasmid system was based on the pUC 8 cloning vector [19]. The first plasmid created was the pPOL-1 208-1. It contained a monomer of the *Lytechinus variegatus* 5S ribosomal DNA sequence [20] and a 94 basepair (bp) (-75 to +19) portion of the RNA

polymerase I promoter sequence (derived from *Acanthamoeba castellanii* [21]). The individual inserts were amplified by polymerase chain reaction (PCR) and provided with the proper restriction enzyme recognition sequence terminations. The individual fragments were ligated together and inserted into the Pst I restriction site of pUC 8.

3.2.1 Amplification of the RNA polymerase I promoter sequence

The plasmid vector pEBH10 [22] (kindly provided by Marvin Paule of Colorado State University) harboring a 200 bp fragment of the RNA polymerase I promoter from *Acanthamoeba castellanii* was amplified by the small scale plasmid preparation method and the DNA was used as template for a PCR reaction.

The PCR primer sequences were designed to selectively amplify a 94 bp (-75 to +19) portion of the RNA polymerase I promoter sequence. According to footprinting experiments performed on the promoter [23] sufficient promoter sequence is included in this fragment to allow specific binding of the transcription initiation complex and necessary factors for transcription.

The PCR primer sequences contain a Pst I restriction site on the 5' terminus of the promoter and a Xba I restriction site on the 3' terminus. Furthermore, a 4 bp extension at the ends of the primer sequences was included to assure a satisfactory digestion of the PCR products with restriction endonucleases. If a 4 bp extension is provided, the nuclease might have a better binding substrate and digestion might be more efficient. The sequence of the promoter fragment and of the primer sequences for the PCR reaction are shown in Figure 3.2.1.

The promoter was amplified according to the PCR procedure outlined in the Materials and Methods section (page 84) with following modification: the elongation cycle was shortened to 2 minutes, because of the short sequence. Also, the amount of TAQ polymerase was varied between 2 and 5 units per reaction. The PCR product was screened by 1% agarose gel

```
                                      Pst-1
                                        |
Primer-1:        5'-AACTCTGCAGGATCCCGGCAAACCGCAAA-3'
                             ||||||||||||||||||||
                 3'-CCCGGCCTTTTTAGGGCCGTTTGGCGTTTGGCCGGCCTCGTGAAAA...
                 5'-GGGCCGGAAAAATCCCGGCAAACCGCAAACCGGCCGGAGCACTTTT...
Promoter
sequence         ...AGGCTTTGATATATATTTCCCTGCCCAGGCCGGCCTTTTTTTTTTAT-5
                 ...TCCGAAAGTATATATAAAGGGACGGGTCCGGCCGGAAAAAAAAAAATA-3'
                              ||||||||||||||||||
Primer-2:                  3'-TTTCCCTGCCCAGGCCGGCAGATCTTCAA-5'
                                                      |
                                                    Xba-1
```
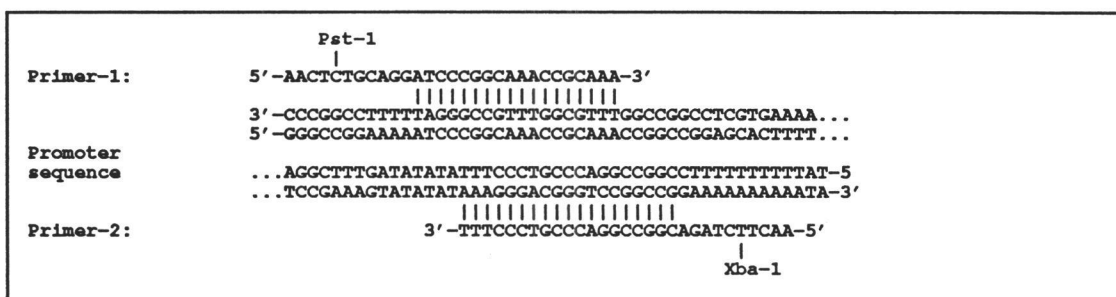
Figure 3.2.1: Primer sequences for the RNA polymerase I promoter sequence. The sense (+) strand is shown on the bottom.

electrophoresis (Figure 3.2.2), isolated by the "Geneclean" method and cloned into pUC 19 after

restriction with Pst 1 and Xba 1 for future use. The complete sequence of the amplified product

is shown in Figure 3.2.3.

```
RNA Polymerase I promoter sequence
after PCR:

      Pst I
        |
5'-AACTGCAGGATCCCGGCAAACCGCAAACCGGCCGGAGCAC

TTTTCTGGCACCTAAACTGGTCGGACCGTCCGAAAGTATATAT

AAAGGGACGGGTCCGGCCGTCTAGATCAA-3'
                         |
                       Xba I
```
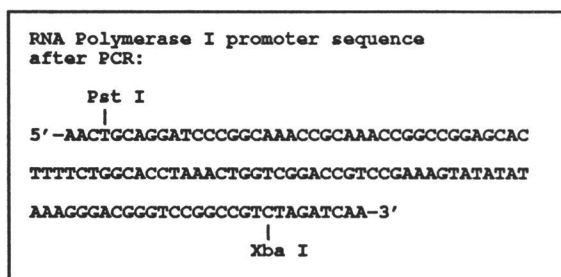
Figure 3.2.3: Complete sequence for PCR amplified polymerase I promoter sequence.

3.2.2 Amplification of the 5S ribosomal sea urchin sequence

The nucleosome positioning sequence (5S ribosomal sequence from sea urchin) was obtained from our own stock of pAT153, harboring a 208 bp 12-mer repeat of the sequence. In addition, two shorter versions of the same sequence were amplified, the 190 and 172 bp repeat length. These sequences had 18 bp and 36 bp, respectively, deleted from the 3' end of the 5S sequence. A small scale plasmid amplification was prepared and the purified DNA

Figure 3.2.2: PCR reaction products for the promoter sequence.

used as template for the PCR reaction. The PCR primer sequences were designed to contain a Xba I restriction site on the 5' terminus of the 5S sequence and to incorporate a Pst I restriction site on the 3' terminus. Again, a 4 bp extension at the ends of the primer sequences was included to assure a satisfactory digestion of the PCR products with restriction endonucleases.

Two different sets of primers were made for the 5S sequence, the first set incorporated the Xba I site on the 5' end and the Pst I site on the 3' end of the sequence (primer set 4/7). In the other set, the Pst I site was switched to the 5' end and the Xba I site was incorporated on the 3' end (primer set 6/8). Two different sets of primers were used to allow for the construction of ligations with either orientation of the 5S sequence with respect to the RNA polymerase I promoter site (Figure 3.2.4):

```
                               Hha I
                                |
5S sequence:               5'-CGCATCTCGGGGAATTCGGTATTCCCA...
                           3'-GCGTAGAGCCCCTTAAGCCATAAGGGT...
                             ||||||||||
Primer-8:         5'-AACTCTGCAGGCGCATCTCGG (Pst I)
Primer-7:         5'-AACTTCTAGAGCGCATCTCGG (Xba I)


Primer-6:            3'-CCAGGACCGGTGCCCACGCGAGATCTCAA-5'  (Xba I)  KVH 6
Primer-4:            3'-CCAGGACCGGTGCCCACGCGACGTCTCAA-5'  (Pst I)  KVH 4
                      ||||||||||||||||||||||
5S sequence:      ...GCGTTGGGTCCTGGCCACGGGTGCG-3'
                  ...CGCAACCCAGGACCGGTGCCCACGC-5'
                                      |
                                    Hha I
```
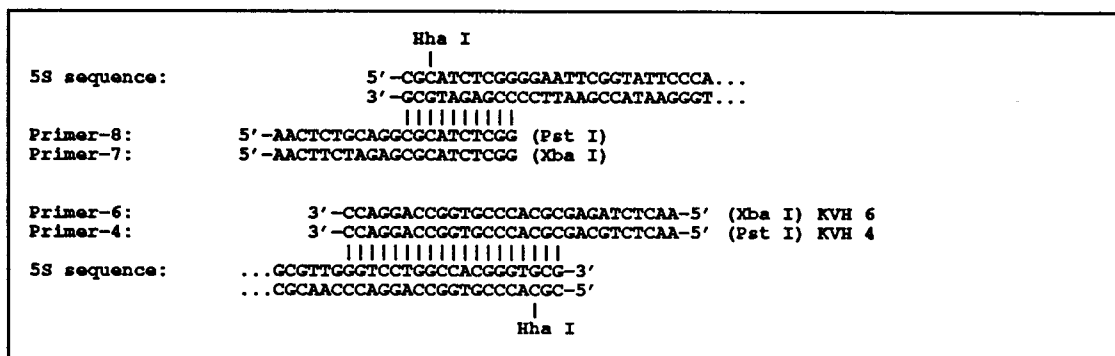
Figure 3.2.4: PCR primer sequences for the ribosomal 5S sequence from sea urchin. The sense (+) strand is shown on top.

Initially, it was attempted to amplify a 12-mer repeated version of the 5S sequence. However, amplification of the 5S sequence proved difficult and yields for the full length product were excessively low for three reasons:

(1) During the PCR reaction, extensive secondary structure is formed by the melted strands causing self priming,

(2) a partial binding site for one of the primers exists in the first round of amplification in every repeated unit of the 12-mer, and

(3) in later PCR cycles the amplification of shorter fragments is favored over the longer products. The amplification procedure was modified several times to try to optimize stringency and amplification conditions. The first amplification round yielded only amplification products shorter than 12 repeats (Figure 3.2.5), with a smear of longer products and only the 1, 2, 3, 4, 5 and 6 repeated 5S molecules amplified somewhat visible on the gel.
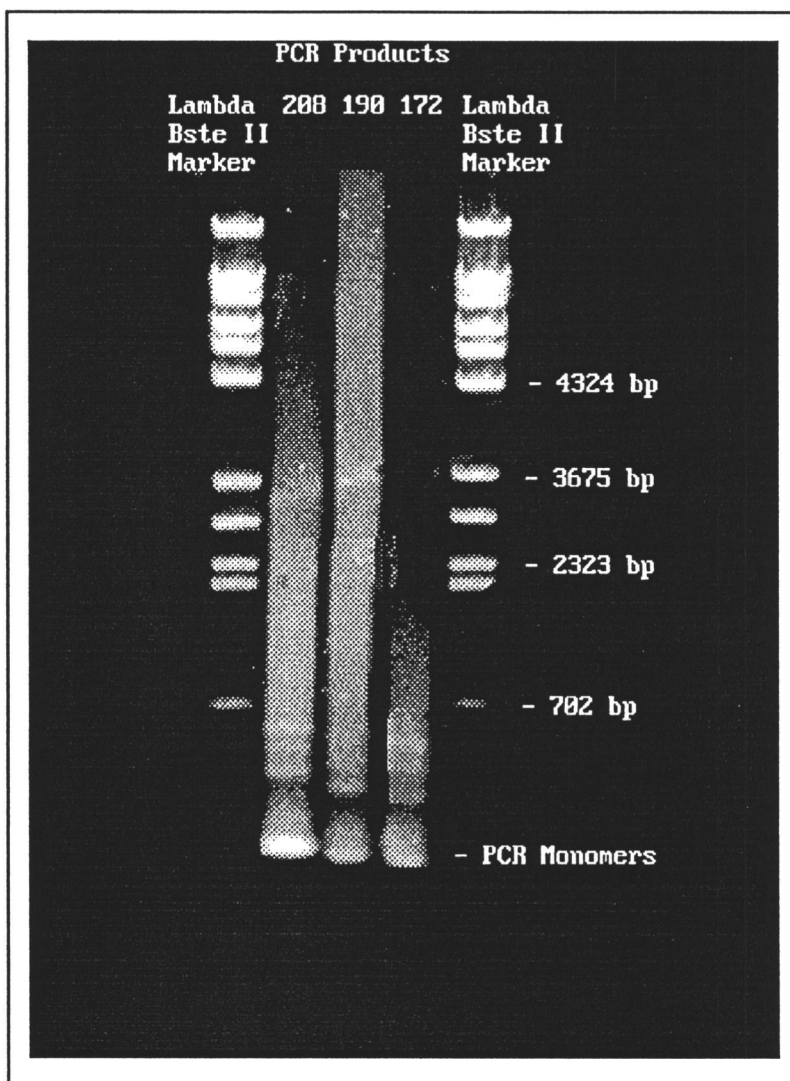
Figure 3.2.5: Amplification of 208-12, 190-12 and 172-12 5S sequences by PCR.

The amplification mixture was screened by 6% polyacrylamide electrophoresis (PAGE) and bands from individual repeats were cut out from the gel, purified and amplified again, but results were not much better. Yields for the purified amplified product were too low to successfully clone the fragments (Figure 3.2.6). However, the single repeat was preferentially amplified, and sufficient DNA could be purified from the polyacrylamide gels to accomplish a

successful clone, after a single repeat was amplified by isolating a monomeric amplification product from a 6% polyacrylamide gel and reamplifying the monomer by itself (Figure 3.2.7).

After purification from polyacrylamide, a portion of the purified product was digested with Pst 1 and Xba 1 and cloned into pUC 19, which has both sites available as unique restriction sites.
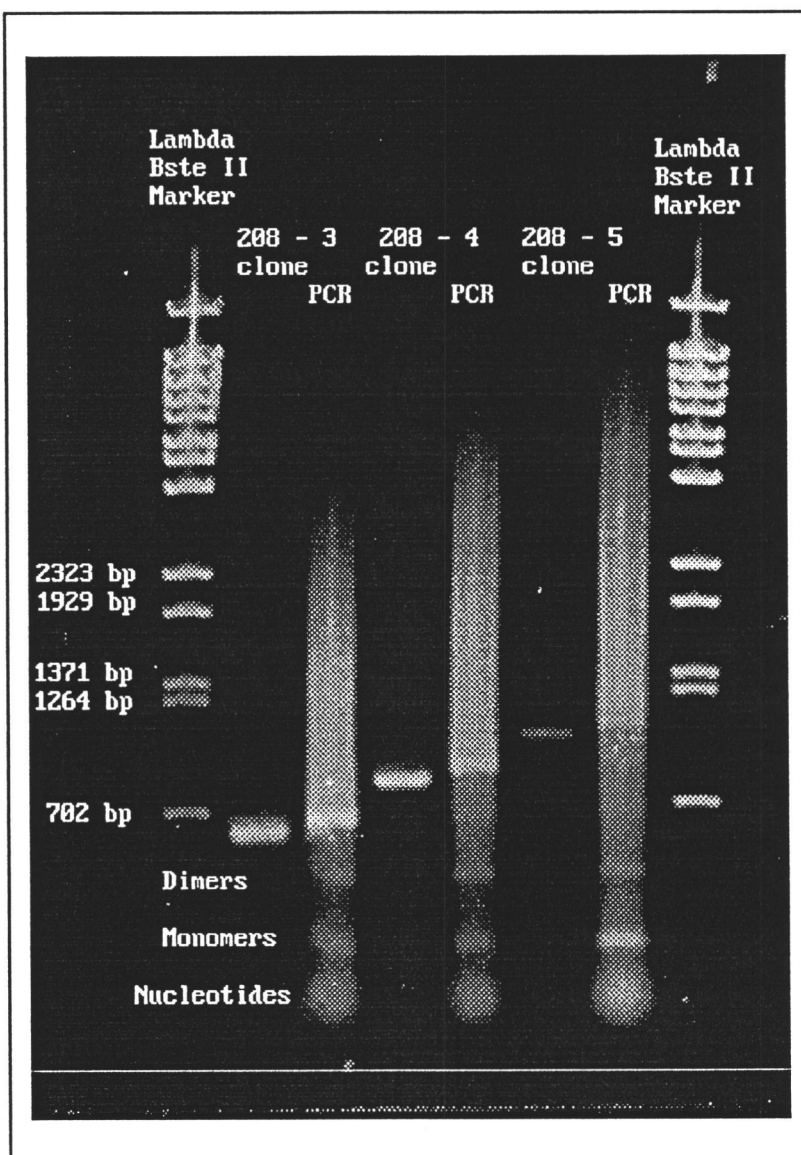
Figure 3.2.6: Amplification of 208-3, 208-4 and 208-5 by PCR.

All further cloning experiments from here on were performed only on the 208 bp length repeat and on the 172 bp length repeat of the 5S sequence. Both orientations of the 5S 208 sequence were cloned, with and without the promoter. The complete sequence for the 208 monomer with some of the adjacent sequence is shown in Figure 3.2.8. The underlined portion constitutes the repeated monomer.

The 208 bp sequence contains on either end a recognition site for the asymmetrically cutting Ava I restriction endonuclease (Figure 3.2.8). By cutting with Ava I, individual repeat monomers can be polymerized in a head to tail fashion with all polymerized repeats oriented in the same direction. This polymerized construct can then be inserted into a vector containing an Ava I recognition site.
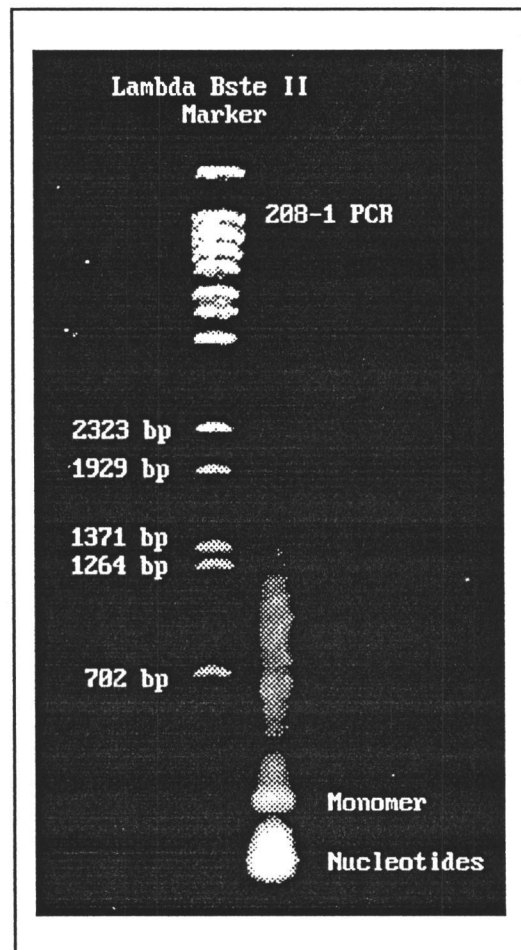
Figure 3.2.7: 208-1 monomer of 5S sea urchin gene amplified with PCR.

```
Underlined = repeated sequence

               Hha I  Ava I            Nci I
                 |      |                 |
        5'-   CATC|TCGG GGAATTCGGTATTCCCAGGCGGTCTCCCATCCAAGTACTAACCG
        3'-GCGTAG AGCC|CCTTAAGCCATAAGGGTCCGCCAGAGGGTAGGTTCATGATTGGC
            ||||||  ||||          | Nci I
  5'-AACTCTGCAGGCGCATC TCGG (Pst I) Primer 8
  5'-AACTTCTAGAGCGCATC TCGG (Xba I) Primer 7


AGCCCTATGCTGCTTGACTTCGGTGATCGGACGAGAACCGGTATATTCAGCATGGTATGGTCGTAGGCTCTTGCTTGAT
TCGGGATACGACGAACTGAAGCCACTAGCCTGCTCTTGGCCATATAAGTCGTACCATACCAGCATCCGAGAACGAACTA

GAAAGTTAAGCTATTTAAAGGGTCAGGGATGTTATGACGTCATCGGCTTATAAATCCCTGGAAGTTATTCGTTGGAATT
CTTTCAATTCGATAAATTTCCCAGTCCCTACAATACTGCAGTAGCCGAATATTTAGGGACCTTCAATAAGCAACCTTAA


           3'-CCAGGACCGGTGCCCACGCGAGATCTCAA-5' (Xba I) Primer 6
           3'-CCAGGACCGGTGCCCACGCGACGTCTCAA-5' (Pst I) Primer 4
              ||||||||||||||||||||
  CC|TCGG GCAGCCGTTGGGTCCTGGCCACGGGTGCG-3'
  GG AGCC|CGTCGCAACCCAGGACCGGTGCCCAC  -5'
  |                                |
  Ava I                            Hha I
```

Figure 3.2.8: Complete sequence for the PCR reaction of the 208 bp sequence repeat. Both primer sets and their position of hybridization are shown.

3.2.3 Ligation of PCR products

The PCR amplified promoter sequence and the two amplified orientations of the 208 bp 5S sequence were ligated into pUC 19 [19] restricted with Xba I and Pst I, both for long term storage as well as for the possibility of generating larger amounts of insert DNA. The individual fragments were sequenced by Walter Lang from our laboratory and the sequences were verified.

After preparing a large scale purification of each plasmid, all plasmids were linearized with Xba I. The linearized plasmids containing the 208 bp sequences (both directions) were separately mixed with equal molar amounts of the linearized plasmid containing the promoter sequence. A ligation was performed and the ligation mixture was digested with Pst I. There are several possible ligation products which can be readily distinguished by screening in 6% polyacrylamide electrophoresis. This electrophoresis system sufficiently resolves the size difference and allows purification of the promoter/208 sequences from the gel. A gel resolving the ligation/restriction products is shown in Figure 3.2.9, and Figure 3.2.10 shows a graphic

representation of possible ligation/restriction products.

The band corresponding to the promoter/208-5S ligation sequence was isolated from the gel and cloned into pUC 8 after digesting the vector with Pst I. Monomer clones were isolated and the presence of inserts was verified by digestion with Pst I and Xba I.
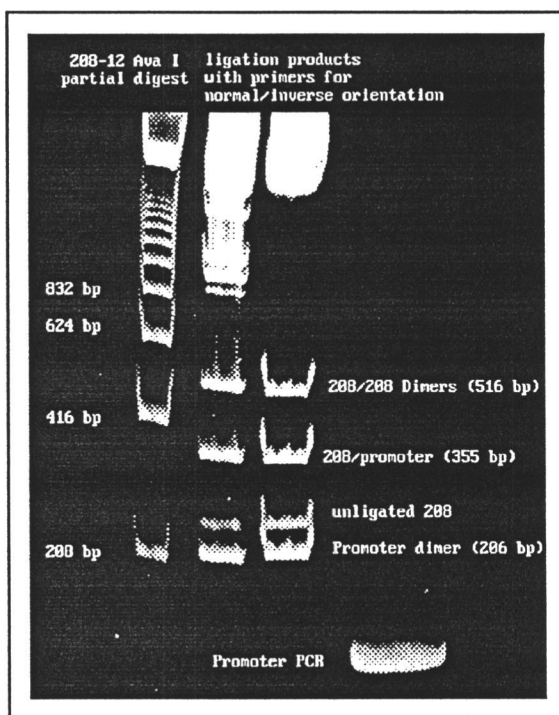
Figure 3.2.9: Ligation products for the ligation of promoter and 208 bp monomer sequence.
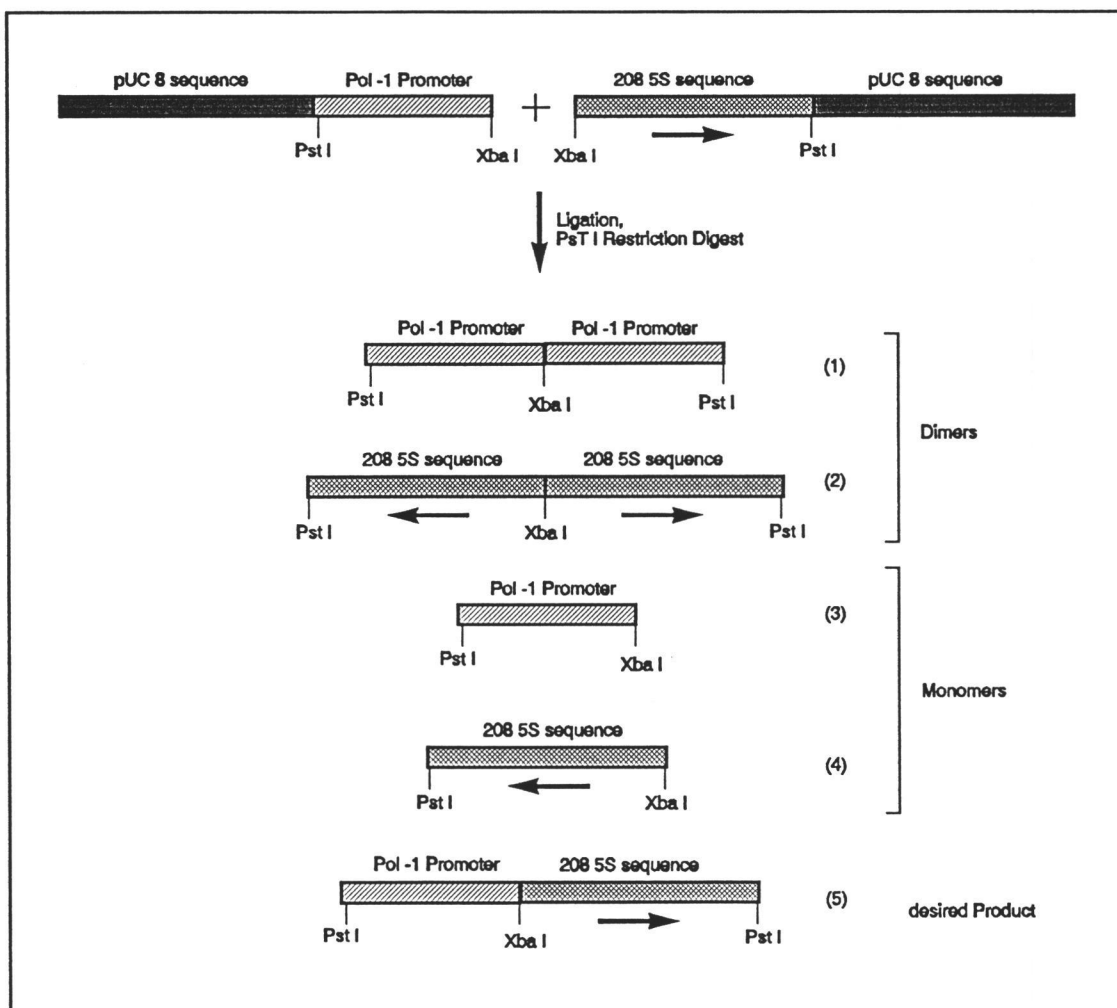
Figure 3.2.10: Possible reaction products of ligation for promoter with 208 bp 5S monomer sequence. Only the ligation for one direction of the 208 bp sequence is shown.

The complete sequence of the ligation products for the promoter/208 bp fragment and their restriction chart for several single and double cutting enzymes is shown in Figure 3.2.11, 11-A (primers 4 and 7) and in Figure 3.2.12, 12-A (primers 6 and 8). The sequence for the promoter/172 bp fragment is shown in Figure 3.2.13, 13-A (primers 4 and 7) and Figure 3.2.14, 14-A (primers 6 and 8).

```
Promoter/208 bp sequence (primers 4 and 7)

                EcoRI                    PstI
                  |                        |
   1 ATGACCATGATTACGAATTCGGATCCGTCGACCTGCAGGATCCCGGCAAACCGCAAACCGGCCGGAGCACT  71
                16                       38

                                                      XbaI     HhaI
                                                       |        |
  72 TTTCTGGCACCTAAACTGGTCGGACCGTCCGAAAGTATATATAAAGGGACGGGTCCGGCCGTCTAGAGCGC  142
                                                      134      142

         AvaI  EcoRI
          |     |
 143 ATCTCGGGGAATTCGGTATTCCCAGGCGGTCTCCCATCCAAGTACTAACCGAGCCCTATGCTGCTTGACTT  213
         146   152

 214 CGGTGATCGGACGAGAACCGGTATATTCAGCATGGTATGGTCGTAGGCTCTTGCTTGATGAAAGTTAAGCT  284

                                            XmnI         EcoRI AvaI
                                             |             |    |
 285 ATTTAAAGGGTCAGGGATGTTATGACGTCATCGGCTTATAAATCCCTGGAAGTTATTCGTTGGAATTCCTC  355
                                            338          348  354

                        HhaI PstI
                         |    |
 356 GGGCAGCCGTTGGGTCCCTGGCCACGGGTGCGCTGCAGCCAAGCTTGGCACTGGCC  410
                        387  392
```

Figure 3.2.11: Sequence for the promoter/208 bp sequence (and some flanking pUC 8 sequence for primers 4 and 7.

```
Restriction map for promoter/208 bp sequence (primers 4 and 7)

       1            94           188          281          374
       |------------|------------|------------|------------|
enzyme
AvaI   |                         !                       !  |
EcoRI  | !                       !                     !    |
HhaI   |                         !                        ! |
PstI   |    !                                               ! |
XbaI   |                   !                                |
XmnI   |                         !                          |
```
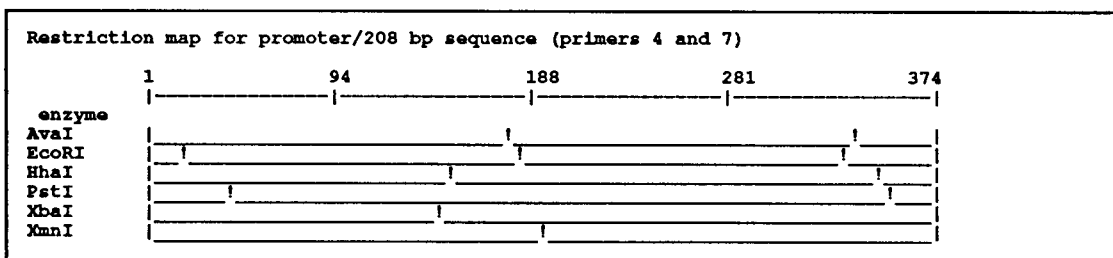
Figure 3.2.11-A: Restriction map for the promoter/208 bp sequence (and some flanking pUC 8 sequence for primers 4 and 7.

```
Promoter/208 bp sequence (primers 6 and 8)

              EcoRI                    PstI
                |                        |
  1 ATGACCATGATTACGAATTCGGATCCGTCGACCTGCAGGATCCCGGCAAACCGCAAACCGGCCGGAGCACT  71
                16                       38

                                                   XbaI     HhaI
                                                     |        |
 72 TTTCTGGCACCTAAACTGGTCGGACCGTCCGAAAGTATATATAAAGGGACGGGTCCGGCCGTCTAGAGCGC 142
                                                    134      142

                   AvaI    EcoRI           XmnI
                     |        |               |
143 ACCCGTGGCCAGGACCCAACGCTGCCCGAGGAATTCCAACGAATAACTTCCAGGGATTTATAAGCCGATGA 213
                    168      174             188

214 CGTCATAACATCCCTGACCCTTTAAATAGCTTAACTTTCATCAAGCAAGAGCCTACGACCATACCATGCTG 284

                                        EcoRI AvaI       HhaI
                                          |    |           |
285 AATATACCGGTTCTCGTCCGATCACCGAAGTCAAGCAGCATAGGGCTCGAATTCCCCGAGATGCGCCTGCA 355
                                         334  340         350
    PstI
      |
356 GCCAAGCTTGGCACTGGCC 374
```

Figure 3.2.12: Sequence for the promoter/208 bp sequuence (and some flanking pUC 8 sequence) for primers 6 and 8.

```
Restriction map for promoter/208 bp sequence (primers 6 and 8)

          1          94          188         281         374
          I-----------I-----------I-----------I-----------I
  enzyme
AvaI      I------------------------'---------------------'---I
EcoRI     I--'------------------------'------------------'----I
HhaI      I-------------------'-----------------------------'-I
PstI      I-----'----------------------------------------'---I
XbaI      I-------------------'------------------------------I
XmnI      I--------------------------'----------------------I
```
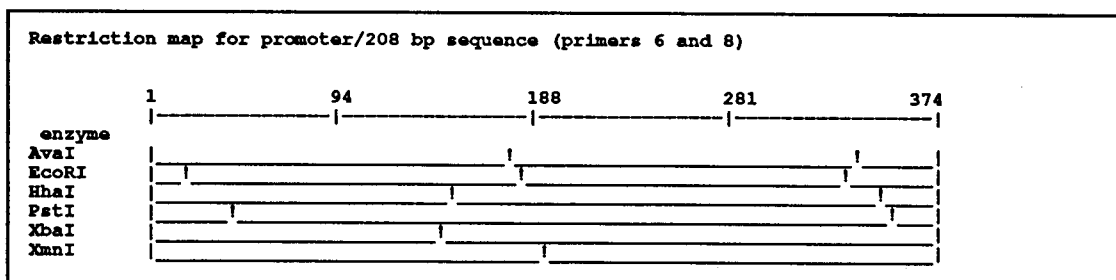
Figure 3.2.12-A: Restriction map for the promoter/208 bp sequence (and some flanking pUC 8 sequence) for primers 6 and 8.

```
Promoter/172 bp sequence (primers 4 and 7)

                    EcoRI                    PstI
                    |                        |
    1 ATGACCATGATTACGAATTCGGATCCGTCGACCTGCAGGATCCCGGCAAACCGCAAACCGGCCGGAGCACT  71
                    16                       38

                                                        XbaI      HhaI
                                                        |         |
   72 TTTCTGGCACCTAAACTGGTCGGACCGTCCGAAAGTATATATAAAGGGACGGGTCCGGCCGTCTAGAGCGC 142
                                                        134       142

        AvaI  EcoRI
        |     |
  143 ATCTCGGGGAATTCGAGCCCTATGCTGCTTGACTTCGGTGATCGGACGAGAACCGGTATATTCAGCATGGT 213
        146   152

  214 ATGGTCGTAGGCTCTTGCTTGATGAAAGTTAAGCTATTTAAAGGGTCAGGGATGTTATGACGTCATCGGCT 284

                  XmnI        EcoRI AvaI                              HhaI
                  |           |     |                                 |
  285 TATAAATCCCTGGAAGTTATTCGTTGGAATTCCTCGGGCAGCCGTTGGGTCCCTGGCCACGGGTGCGCTGCA 355
                  302         312   318                               351
        PstI
        |
  356 GCCAAGCTTGGCACTGGCC 374
        356
```

Figure 3.2.13: Sequence for the promoter/172 bp sequence (and some flanking pUC 8 sequence) for primers 4 and 7.

```
Restriction map for promoter/172 bp sequence (primers 4 and 7)

          1           94            188           281           374
          |-----------|-------------|-------------|-------------|
   enzyme
   AvaI   |                  !                         !        |
   EcoRI  |  !                      !                      !     |
   HhaI   |                   !                            !     |
   PstI   |      !                                           !   |
   XbaI   |                         !                             |
   XmnI   |                         !                      !      |
```

Figure 3.2.13-A: Restriction map for the promoter/172 bp sequence (and some flanking pUC 8 sequence) for primers 4 and 7.

```
Promoter/172 bp sequence (primers 6 and 8)

                EcoRI                    PstI
                 |                        |
  1 ATGACCATGATTACGAATTCGGATCCGTCGACCTGCAGGATCCCGGCAAACCGCAAACCGGCCGGAGCACT  71
                16                       38

                                              XbaI      HhaI
                                               |         |
 72 TTTCTGGCACCTAAACTGGTCGGACCGTCCGAAAGTATATATAAAGGGACGGGTCCGGCCGTCTAGAGCGC 142
                                              134       142

                   AvaI    EcoRI          XmnI
                    |        |             |
143 ACCCGTGGCCAGGACCCAACGCTGCCCGAGGAATTCCAACGAATAACTTCCAGGGATTTATAAGCCGATGA 213
                   168      174          188

214 CGTCATAACATCCCTGACCCTTTAAATAGCTTAACTTTCATCAAGCAAGAGCCTACGACCATACCATGCTG 284

                                         EcoRI AvaI      HhaI
                                           |    |         |
285 AATATACCGGTTCTCGTCCGATCACCGAAGTCAAGCAGCATAGGGCTCGAATTCCCCGAGATGCGCCTGCA 355
                                          334  340       350
    PstI
     |
356 GCCAAGCTTGGCACTGGCC 374
    356
```

Figure 3.2.14: Sequence for the promoter/172 bp sequence (and some flanking pUC 8 sequence) for primers 6 and 8.

```
Restriction map for promoter/172 bp sequence (primers 6 and 8)

        1          94           188          281          374
        |----------|------------|------------|------------|
 enzyme
 AvaI   |          |            |            |          | |
 EcoRI  | |                     |                     |    |
 HhaI   |   |                |                      |      |
 PstI   |     |                                      |  |  |
 XbaI   |                     |                            |
 XmnI   |                  |            |                   |
```
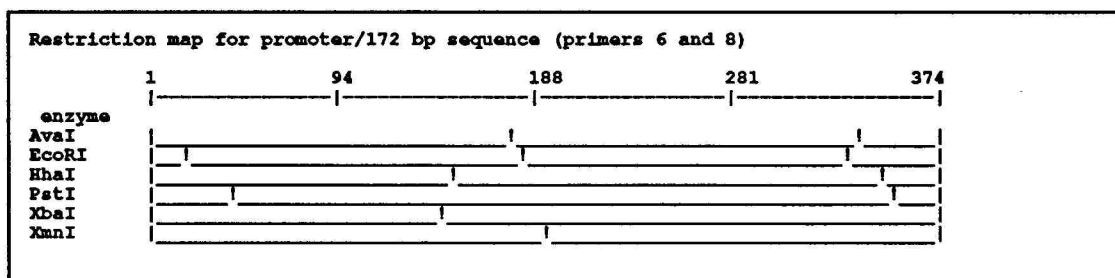
Figure 3.2.14-A: Restriction map for the promoter/172 bp sequence (and some flanking pUC 8 sequence) for primers 6 and 8.

Ligation products were verified by (1) repeating the PCR reaction with the cloned product as template. Both orientations were checked for the 208 sequence and by restriction analysis of the cloned product. The results of the PCR reaction are shown in Figure 3.2.15, and the restriction digests are shown in Figure 3.2.16 (208 bp) and 3.2.17 (172 bp).
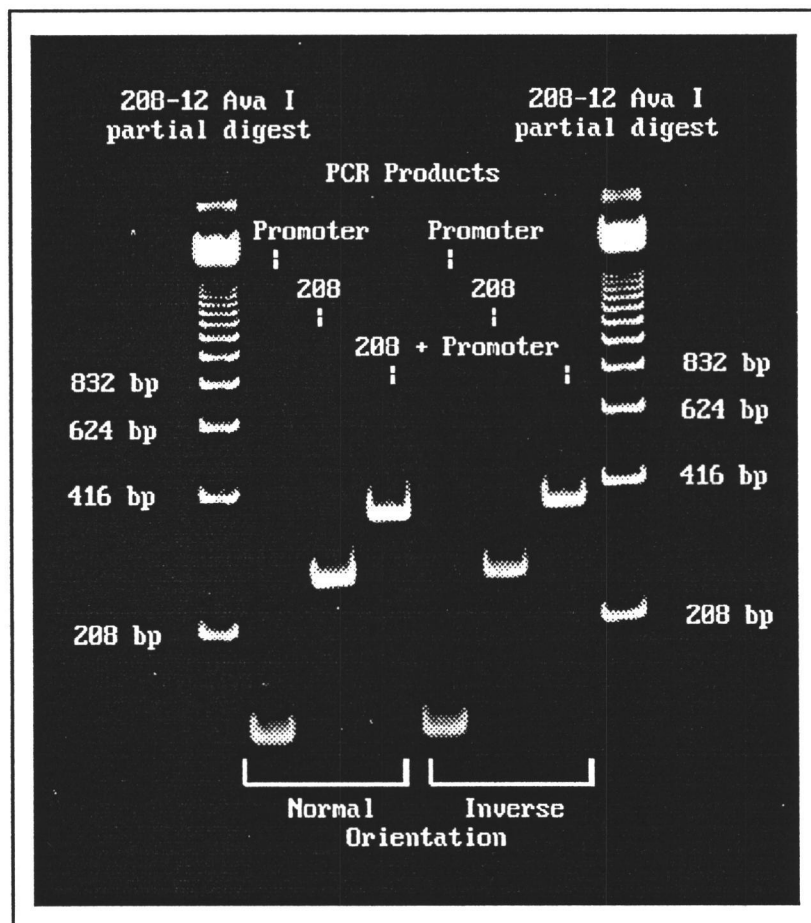
Figure 3.2.15: Products of PCR reaction with the promoter/208 bp sequence with primers 1+2 (2), 4+7 (3), 1+4 (4), 1+2 (5), 6+8 (6), 1+8 (7).
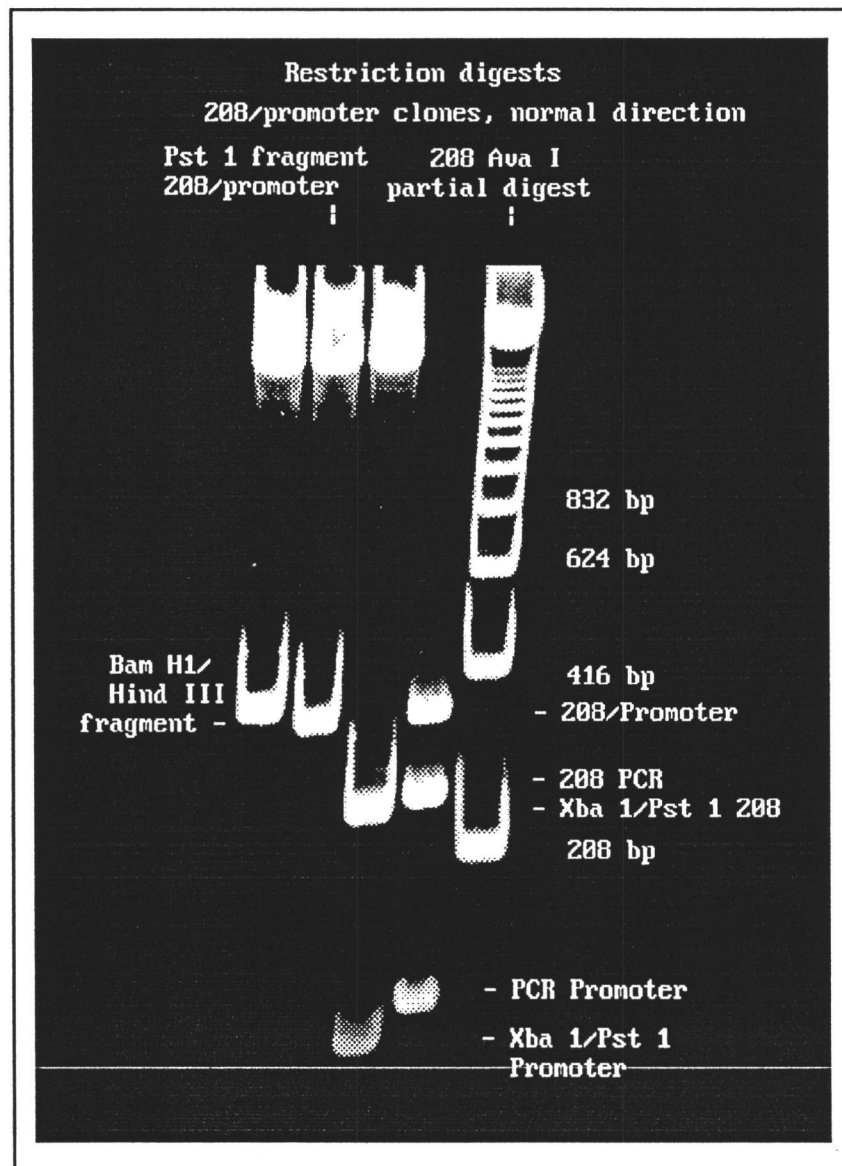
Figure 3.2.16: Restriction digest for the monomer plasmid pPOL-1 208-1 (only results for primers 6 and 8 are shown).
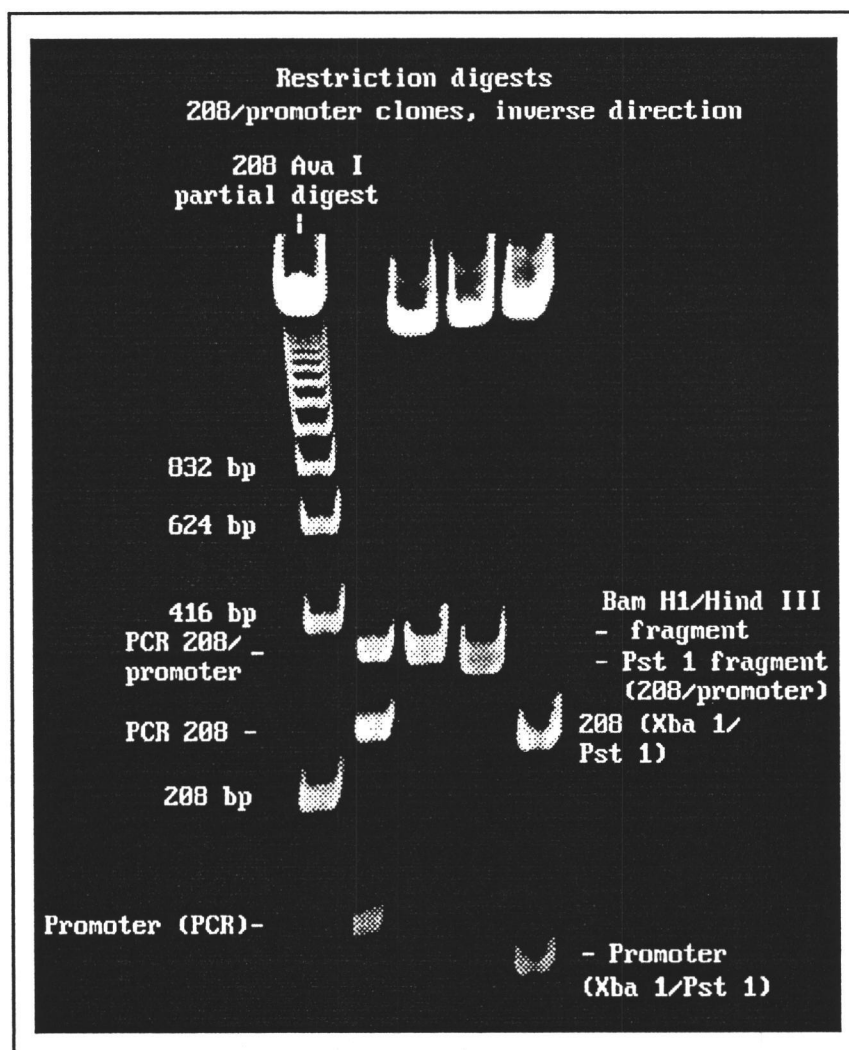
Figure 3.2.17: Restriction digest for monomer plasmid pPOL-1 172-1 (only results for primers 4 and 7 are shown).

3.3 Generating 208 bp/Ava I polymers

The promoter/monomer plasmid was amplified and the monomer was excised with Ava-1, an asymmetric restriction enzyme which cuts once in each repeat and at the 3'-end of the construct. Larger amounts of repeated Ava I units were prepared by partially digesting a pAT153 plasmid containing the 12-mer 208 bp repeat with Ava I.

The repeated units were purified away from the plasmid DNA by high performance electro elution chromatography on a 5 cm column on an Applied Biosystems HPEC machine. The purified 208 bp repeats were polymerized by ligating the mixture of different repeat lengths in one reaction. The asymmetric Ava I site guaranteed head to tail ligation.

A mixture of longer repeat lengths was obtained (Figure 3.3.1) and screened by 1% agarose electrophoresis. Bands corresponding to

Figure 3.3.1: Partial digestion of pAT153 with Ava I and ligation of purified Ava I units.

fragments of length 12 and above units were isolated from the gel and purified by the "Geneclean" method. The promoter/208 bp sequence vector was amplified and the Ava I fragment excised. The vector was now linearized with two Ava I terminals and ready for ligation with the longer fragments.
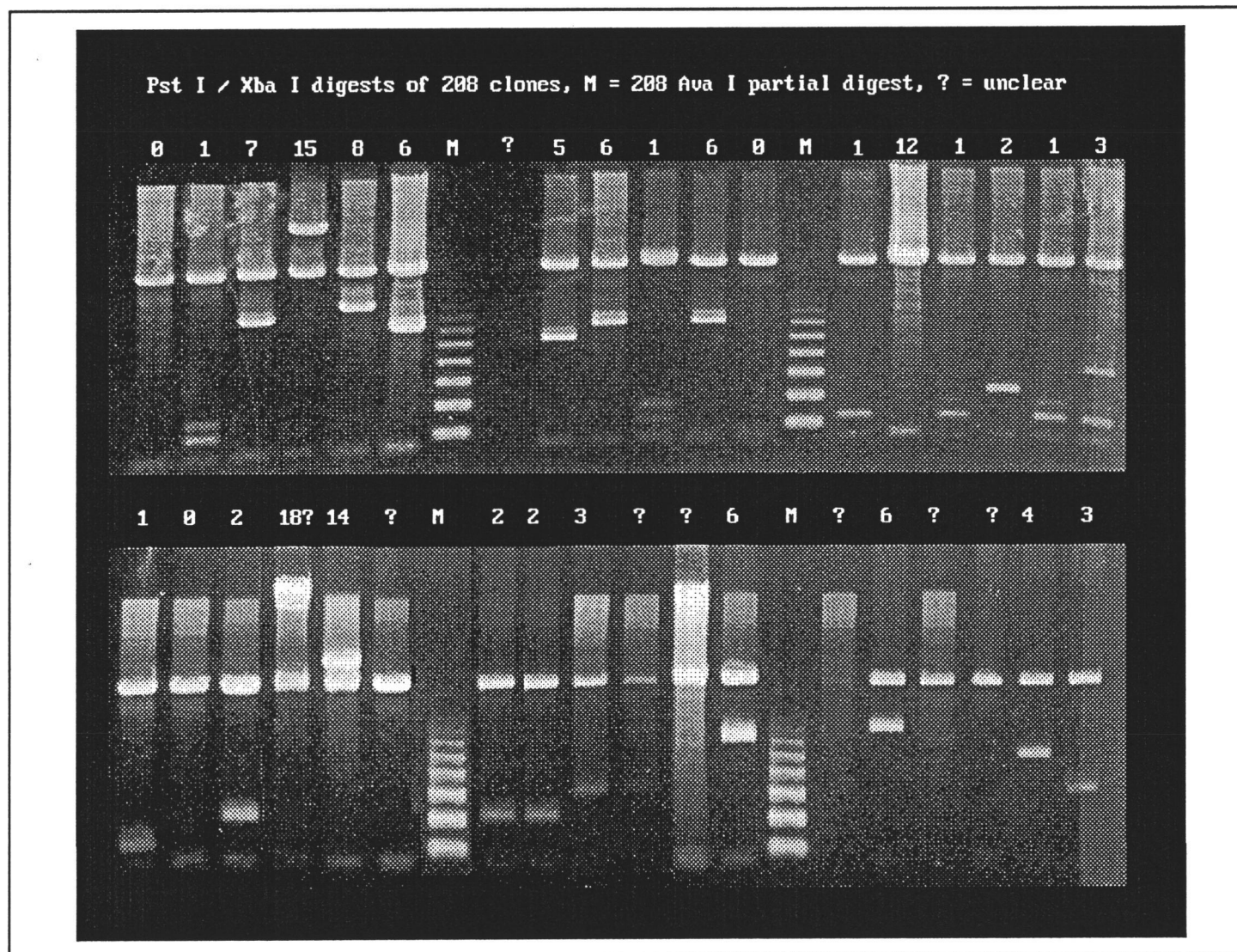
Figure 3.3.2: Restriction digests (Xba I/Pst I) of clones containing the promoter and various repeat numbers of the 208 bp sequence.

The purified fragments 12 units and longer were added to the Ava-1 treated vector in a 100:1 ratio and ligated. The clones were screened for the number of repeats by restriction analysis with Pst I and Xba I. Repeat numbers between 1 and approximately 30 repeats were found (Figure 3.3.2). The plasmid containing the largest repeat was amplified, purified and partially digested with Ava I. The digest was loaded on a long preparative 1% agarose gel and run for 30 hours with low voltage to separate plasmids containing various numbers of 208 bp repeats (Figure 3.3.3).

Plasmid bands containing 1 or more repeats were carefully cut
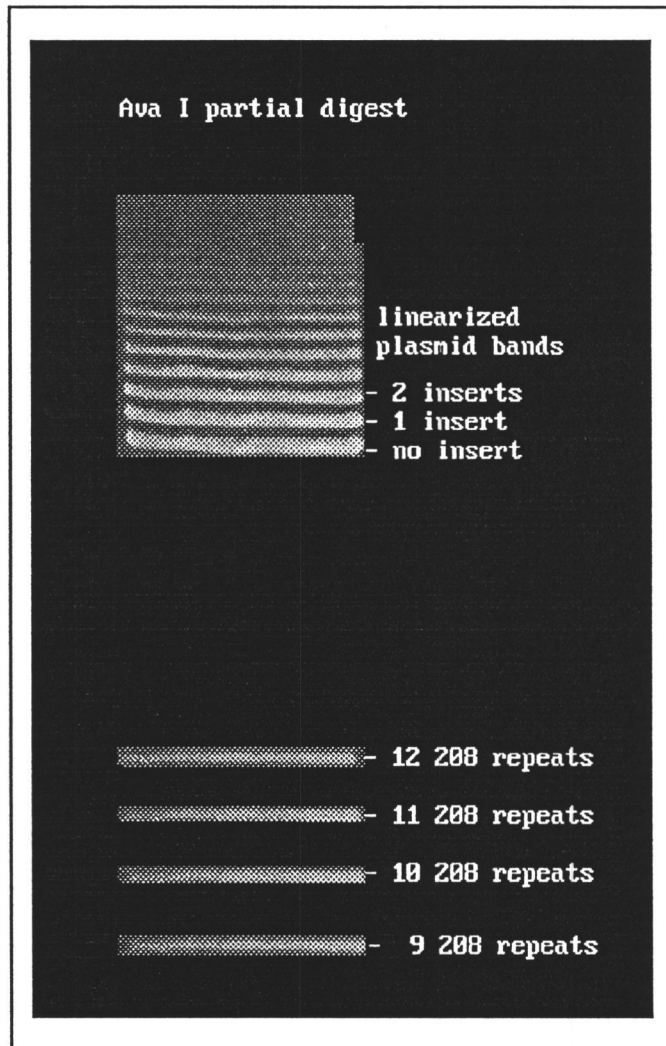
Figure 3.3.3: Partial digest of pAT153 with Ava I. The upper set of bands corresponds to plasmid bands, the lower to insert bands.

out of the gel and purified from the gel by the "Geneclean" method. The purified DNA was religated and transformed into XL1-blue *E. coli* cells. Stocks of clones with repeated 208 bp inserts between 0 and 12 repeats were made. The name for these plasmids is pPOL-1 208-n, where n refers to the number of inserts. A Pst I/Xba I digest of these clones is shown in Figure 3.3.4.
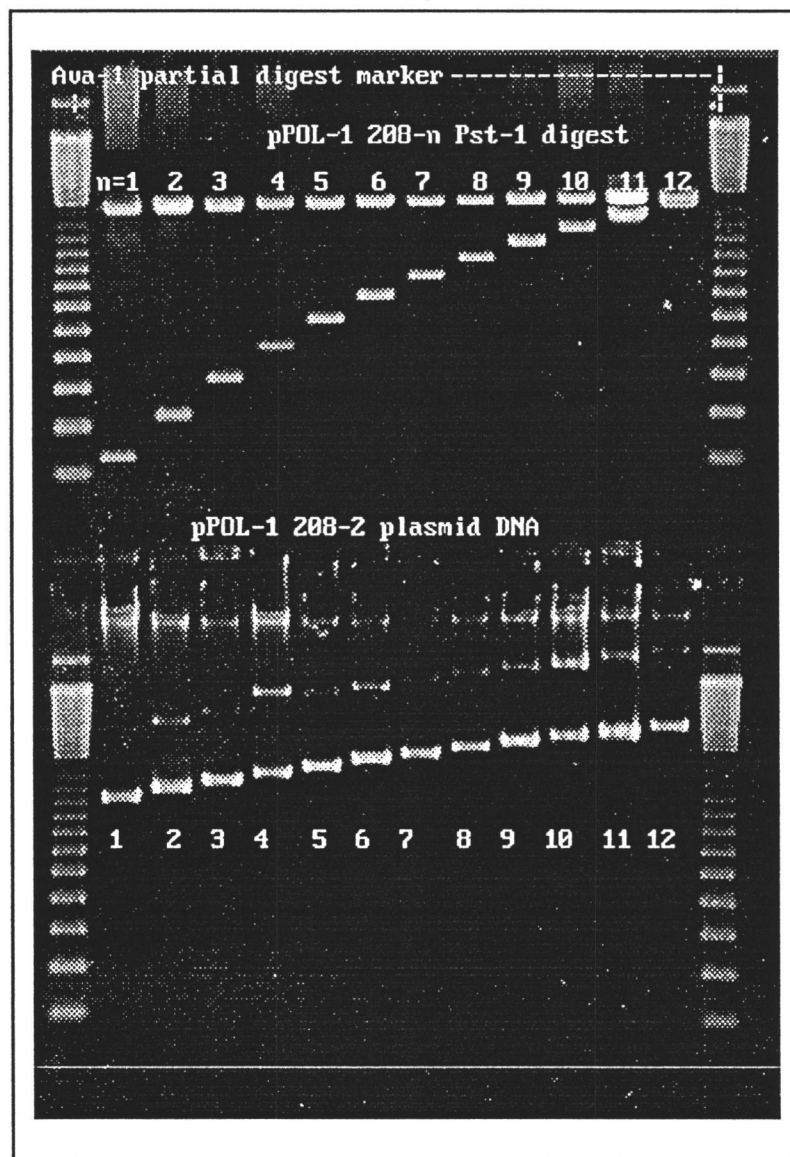
Figure 3.3.4: Series of clones for pPOL-1 208-n, shown are the plasmids (bottom) and Pst I/Xba I digests (top).

3.4 Construction of the miniplasmid pMX

For sedimentation and diffusion studies of reconstitution of circular DNA molecules with nucleosomes it is important that the molecular weight of the reconstituted species not be too large. If the plasmid is large, there are many binding sites for nucleosomes, alas, a high molecular weight. The relative differences in molecular weight between reconstitution with n nucleosomes and n-1 nucleosomes is smaller for large plasmids where n can be quite large (since there are only approximately 140 - 200 bp DNA / nucleosome binding site) and hence more difficult to characterize. Even more difficult is the situation posed in sedimentation velocity or diffusion studies. The percent difference in S_n and S_{n-1}, or D_n and D_{n-1} becomes very small if n is large.

Therefore, for use in these kinds of studies a miniplasmid has been constructed which contains the necessary regulatory sequences for efficient amplification, but no unnecessary extraneous DNA. The sequences necessary for such a plasmid are the replication origin with the RNA primer site and a antibiotic drug resistance gene. The plasmid chosen from which this miniplasmid has been derived was pUC 8. This plasmid contains an origin of replication which has a high copy number and contains the ß-lactamase gene, which confers ampicillin resistance to the host.

Extraneous DNA constituted the inserted cloning region and the ß-galactosidase gene. This region was deleted by digesting with Aat II at position 2622 and Pvu II at position 631. The linearized plasmid was then digested with mung bean nuclease to trim the overhangs and the termini to blunt endings. The DNA was circularized by ligation and transformed into *E.coli* hostcell XL1"blue". A restriction map for the miniplasmid pMX is shown in Figure 3.4.1, the complete sequence for is shown in Figure 3.4.2. The sequence of pUC 8 was obtained from [24], [25], (cloning region from [25]).

This is the PUC8 plasmid sequence
2665 basepairs

Restriction map for pUC 8:

```
                1              667            1333           1999           2665
                |---------------|---------------|---------------|---------------|
        enzyme
        AatII   |---------------------------------------------------------!_|
        PvuI    |---------------!-----------------------------------!--------|
        PvuII   |-----------!--------!---------------------------------------|
```
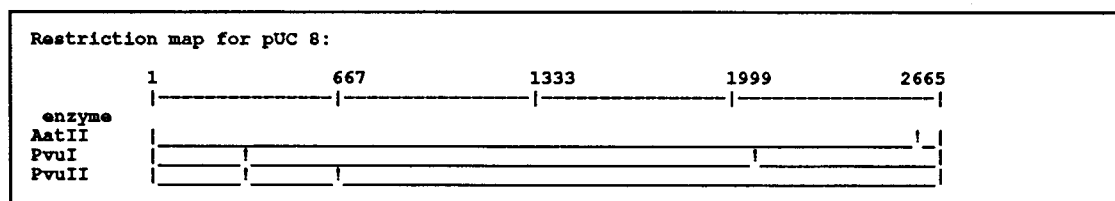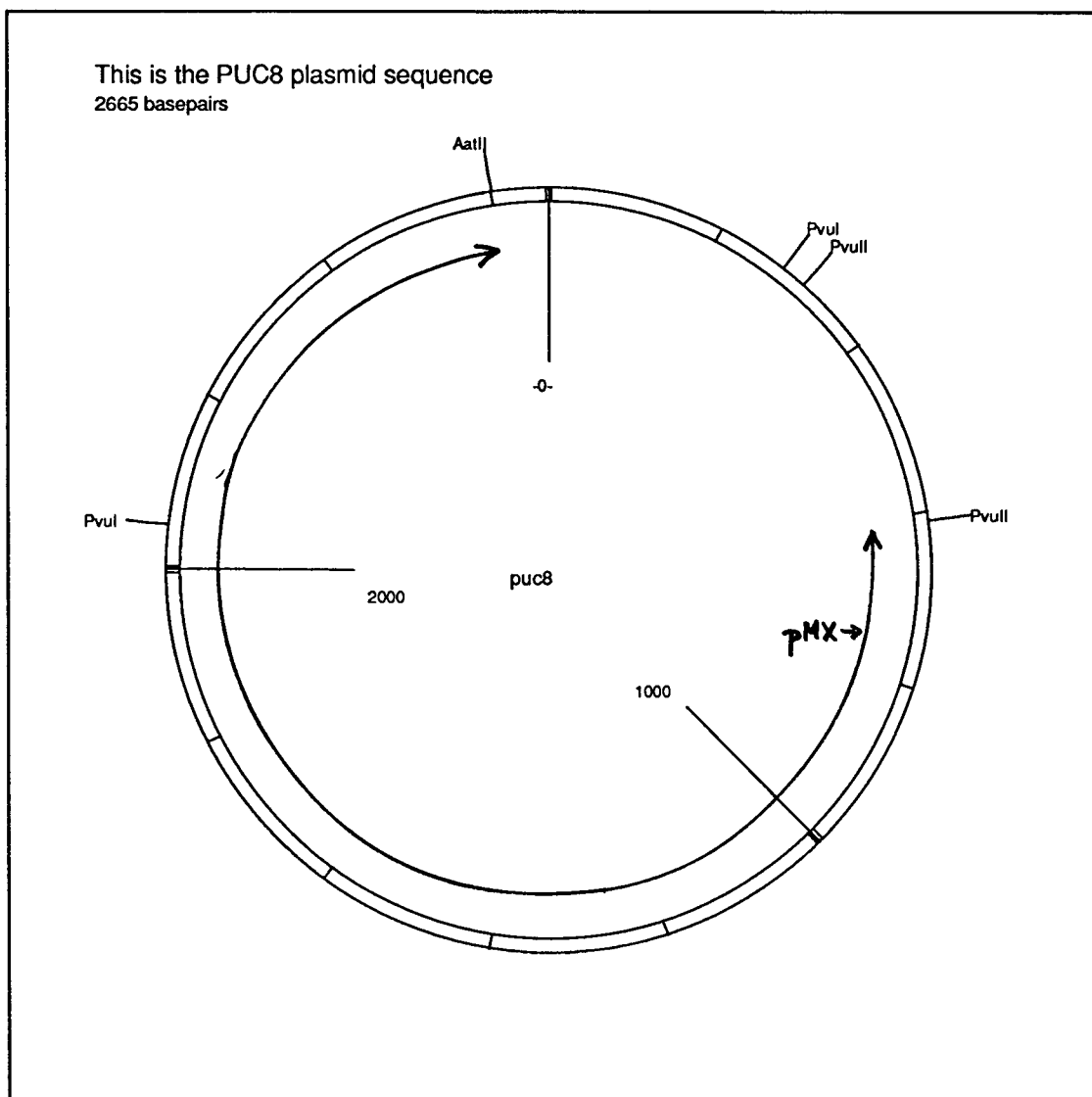
Figure 3.4.1: Restriction map for plasmid pUC 8. Shown are positions for restriction sites of enzymes Aat II, Pvu I and Pvu II.

This is the sequence for the miniplasmid constructed by deleting from Aat II (2622) (GACGT^C) to the second Pvu II site (631) (CAG^CTG).
molecular weight of this strand: 619,436.60 g/mol
molecular weight of complementary strand: 620,281.29 g/mol
molecular weight of entire plasmid: 1,239,717.89 g/mol

```
ctgcattaatgaatcggccaacgcgcggggagaggcggtttgcgtattgggcgct
cttccgcttcctcgctcactgactcgctgcgctcggtcgttcggctgcggcgagc
ggtatcagctcactcaaaggcggtaatacggttatccacagaatcagggataac
gcaggaaagaacatgtgagcaaaaggccagcaaaaggccaggaaccgtaaaaagg
ccgcgttgctggcgtttttccataggctccgcccccctgacgagcatcacaaaa
tcgacgctcaagtcagaggtggcgaaacccgacaggactataaagataccaggcg
tttccccctggaagctccctcgtgcgctctcctgttccgaccctgccgcttaccg
gatacctgtccgcctttctcccttcgggaagcgtggcgctttctcaatgctcacg
ctgtaggtatctcagttcggtgtaggtcgttcgctccaagctgggctgtgtgcac
gaacccccgttcagcccgaccgctgcgccttatccggtaactatcgtcttgagt
ccaacccggtaagacacgacttatcgccactggcagcagccactggtaacaggat
tagcagagcgaggtatgtaggcggtgctacagagttcttgaagtggtggcctaac
tacggctacactagaaggacagtatttggtatctgcgctctgctgaagccagtta
ccttcggaaaaagagttggtagctcttgatccggcaaacaaaccaccgctggtag
cggtggtttttttgtttgcaagcagcagattacgcgcagaaaaaaaggatctcaa
gaagatcctttgatcttttctacggggtctgacgctcagtggaacgaaaactcac
gttaagggattttggtcatgagattatcaaaaaggatcttcacctagatcctttt
aaattaaaaatgaagttttaaatcaatctaaagtatatatgagtaaacttggtct
gacagttaccaatgcttaatcagtgaggcacctatctcagcgatctgtctatttc
gttcatccatagttgcctgactccccgtcgtgtagataactacgatacgggaggg
cttaccatctggccccagtgctgcaatgataccgcgagacccacgctcaccggct
ccagatttatcagcaataaaccagccagccggaagggccgagcgcagaagtggtc
ctgcaactttatccgcctccatccagtctattaattgttgccgggaagctagagt
aagtagttcgccagttaatagtttgcgcaacgttgttgccattgctacaggcatc
gtggtgtcacgctcgtcgtttggtatggcttcattcagctccggttcccaacgat
caaggcgagttacatgatcccccatgttgtgcaaaaaagcggttagctccttcgg
tcctccgatcgttgtcagaagtaagttggccgcagtgttatcactcatggttatg
gcagcactgcataattctcttactgtcatgccatccgtaagatgcttttctgtga
ctggtgagtactcaaccaagtcattctgagaatagtgtatgcggcgaccgagttg
ctcttgcccggcgtcaatacgggataataccgcgccacatagcagaactttaaaa
gtgctcatcattggaaaacgttcttcggggcgaaaactctcaaggatcttaccgc
tgttgagatccagttcgatgtaacccactcgtgcacccaactgatcttcagcatc
ttttactttcaccagcgtttctgggtgagcaaaaacaggaaggcaaaatgccgca
aaaaagggaataagggcgacacggaaatgttgaatactcatactcttccttttc
aatattattgaagcatttatcagggttattgtctcatgagcggatacatatttga
atgtatttagaaaaataaacaaataggggttccgcgcacatttccccgaaaagtg
ccacctg2
```

Figure 3.4.2: Sequence for miniplasmid pMX.

4. SEDIMENTATION AND DIFFUSION STUDIES ON

DNA AND RECONSTITUTED CHROMATIN

A model system was chosen to test the programs UltraScan and LaserScan for reliability and accuracy. Plasmid DNA from the pPOL-1 208-n series and the miniplasmid pMX were analyzed by sedimentation velocity experiments with UltraScan and results were compared to results obtained from diffusion studies with LaserScan.

Plasmid sample	estimated number of negative supercoils	Molecular weight (calculated on basis of nucleotide MW)	number of basepairs
pMX Miniplasmid	12	1239717.89	1973
pPOL-1 208-1	18	1884161.93	3019
pPOL-1 208-6	24	2532765.73	4059
pPOL-1 208-12	32	3311090.29	5307
pPOL-1 208-24	47	4867739.41	7803

Note: The molecular weight is calculated based on nucleotides alone, counterions are not accounted for.

Table 4.1: List of plasmids used in the study. Shown are estimated number of supercoils, molecular weight and number of basepairs for each plasmid. The number of supercoils is estimated based on a superhelical density of 0.06.
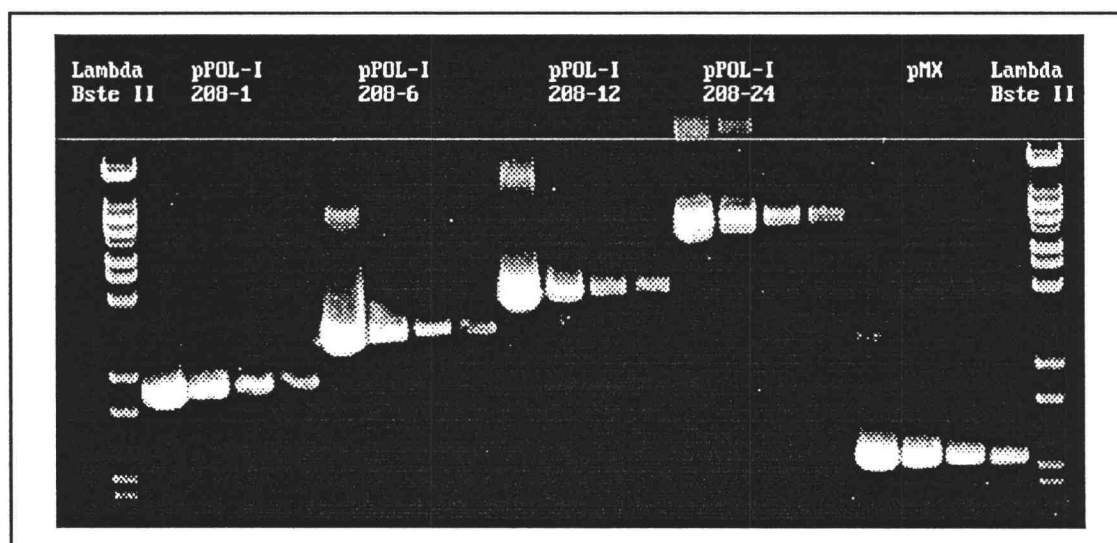
Figure 4.1: Plasmid DNA (1% Agarose gel), purified by CsCl density equilibrium centrifugation. Larger plasmids also show bands corresponding to the linear and open circle (nicked) plasmid conformation.

4.1 Sedimentation studies on plasmid DNA

Sedimentation velocity experiments were performed on all plasmid DNA in TE buffer

(pH 8.0). van Holde - Weischet sedimentation distributions for all plasmids (gel is shown in

Figure 4.1) are shown in Figure 4.1.1.
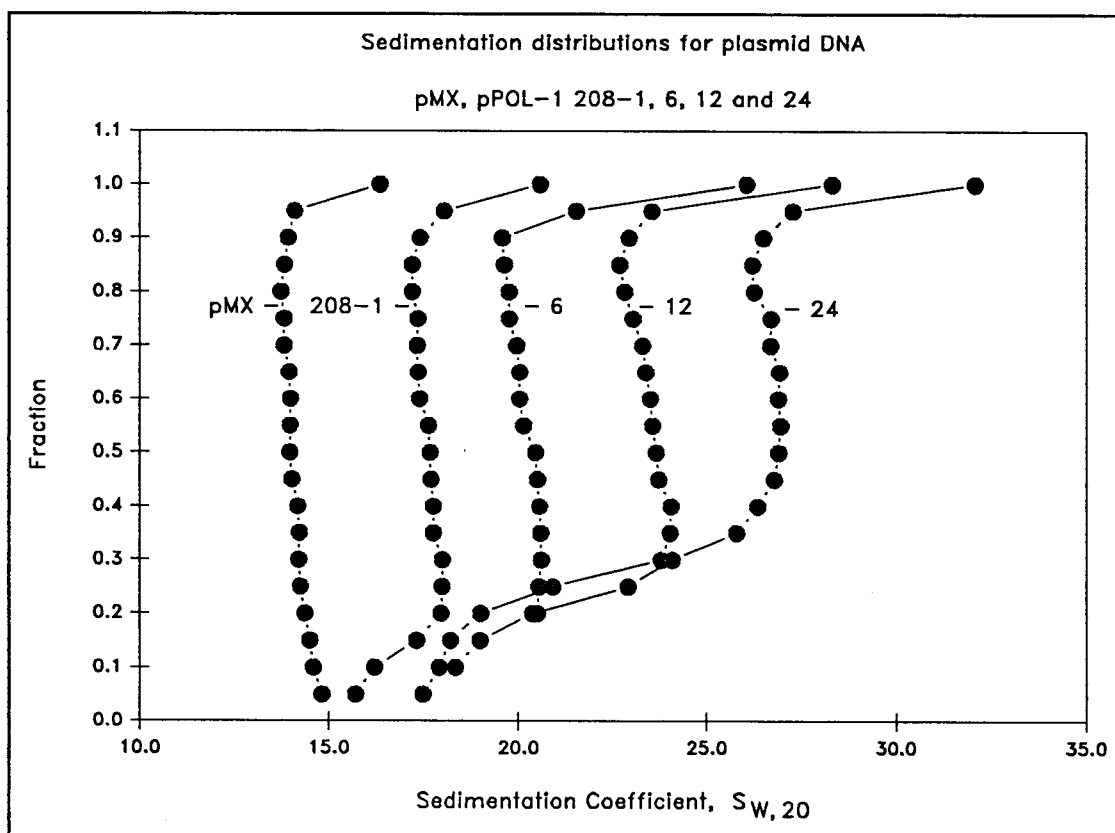
Figure 4.1.1: Temperature corrected van Holde - Weischet sedimentation distributions for all plasmids.

It appears that longer plasmids are less stable in the supercoiled conformation because

slower sedimenting species appear in the distribution. The majority of the plasmid exists in the

supercoiled conformation (see Figure 4.1) and sediments with a higher S-value than the open

circle conformation.

Also, the linear conformation does not exhibit as much change with respect to molecular

weight in sedimentation velocity as the supercoiled species does. A plot of the sedimentation

velocity of the supercoiled species with respect to molecular weight is shown in Figure 4.1.2.
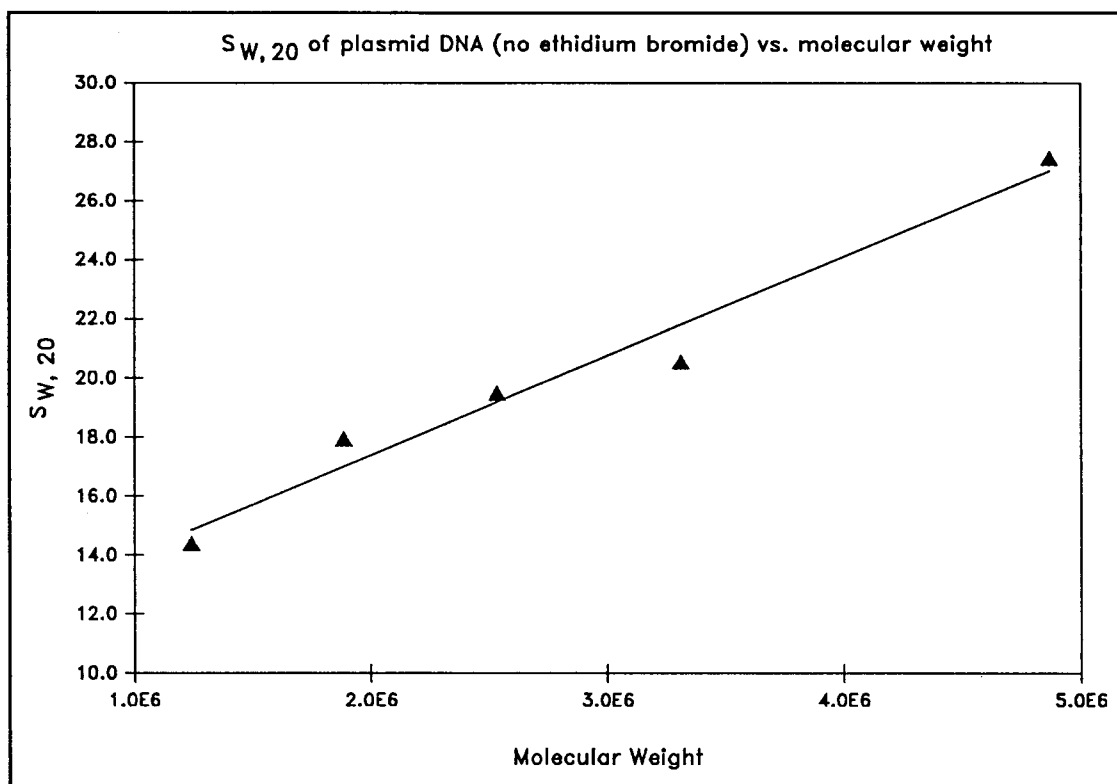
Figure 4.1.2: Plot of sedimentation velocity of supercoiled plasmids versus molecular weight.

Obviously, analysis of sedimentation velocity data by the van Holde - Weischet analysis allows one to differentiate between various conformations of plasmid DNA. This sensitivity to conformation was further demonstrated by measuring the sedimentation coefficient of plasmid pPOL-1 208-24 in a negatively supercoiled conformation (as isolated from E.coli), the same plasmid treated with DNA topoisomerase I for a limited time (generating a number of topoisomers between fully supercoiled and relaxed circle), and the same plasmid digested with Xba I, an enzyme producing a single cut in the plasmid molecule, generating a linear conformation. The result for this experiment is shown in Figure 4.1.3.

The relaxed circular conformation is visible in both the supercoiled sample (to a lesser degree), as well as in the sample treated with topoisomerase I. An interesting feature of the van Holde - Weischet analysis is the sedimentation behavior of plasmid DNA in its dependence on
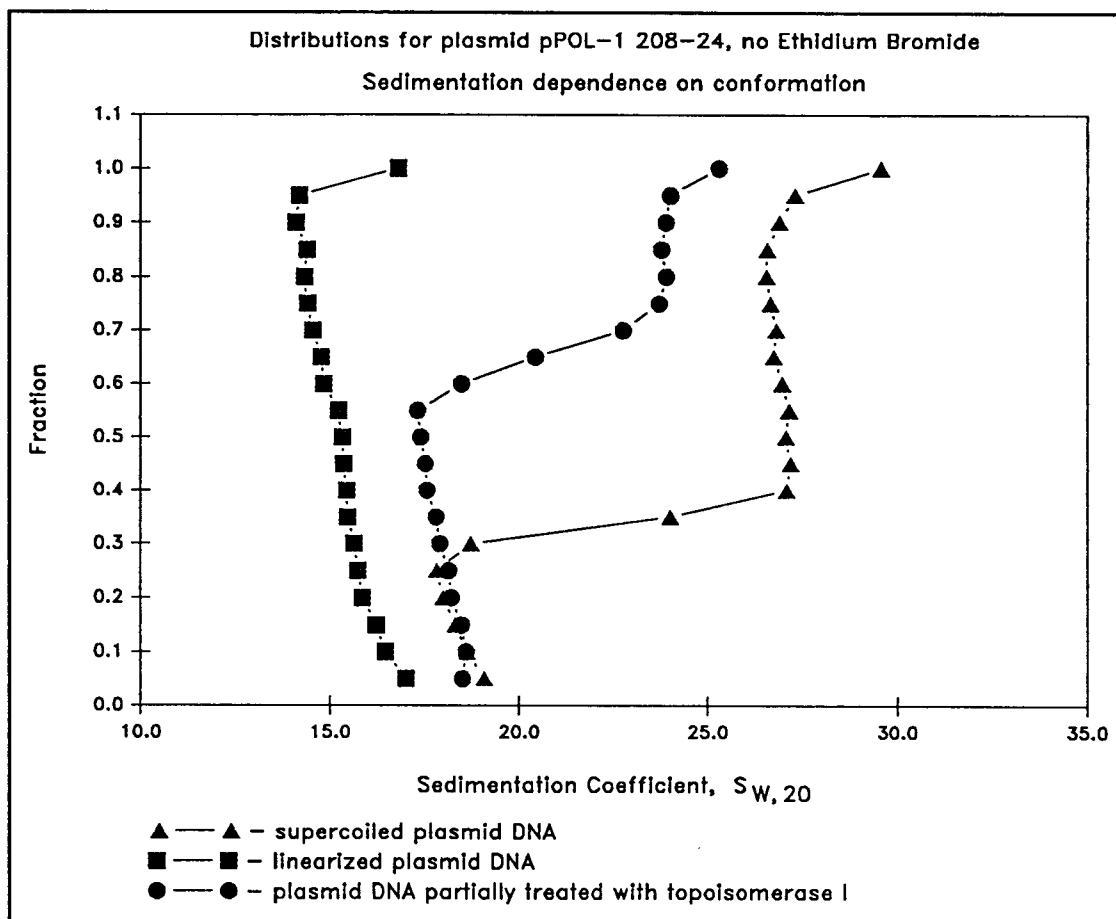
Figure 4.1.3: Sedimentation pattern for pPOL-1 208-24 in different conformations.

on concentration. Higher plasmid concentrations are sedimenting slower, as can be seen when the higher fraction numbers are compared with the lower fraction ranges in graphs 4.1.1 and 4.1.3. Lower fraction numbers correspond to regions in the cell where concentration is lower than in the range for larger fraction numbers.

Extrapolating to zero concentration presumably will result in the true sedimentation coefficient S_0 at which the sample would be sedimenting if no concentration dependence existed. In effect, this provides us with a method for correcting for nonideality as it depends on concentration. The "backward leaning" of sedimentation distributions is thus a feature indicating concentration effects. An extrapolation to zero concentration for plasmid pMX DNA is shown in Figure 4.1.4. It is not clear if a linear extrapolation is the best model for this extrapolation, since

at lower concentration values the datapoints exhibit a slight upward trend. More detailed studies will be necessary to determine the proper model.
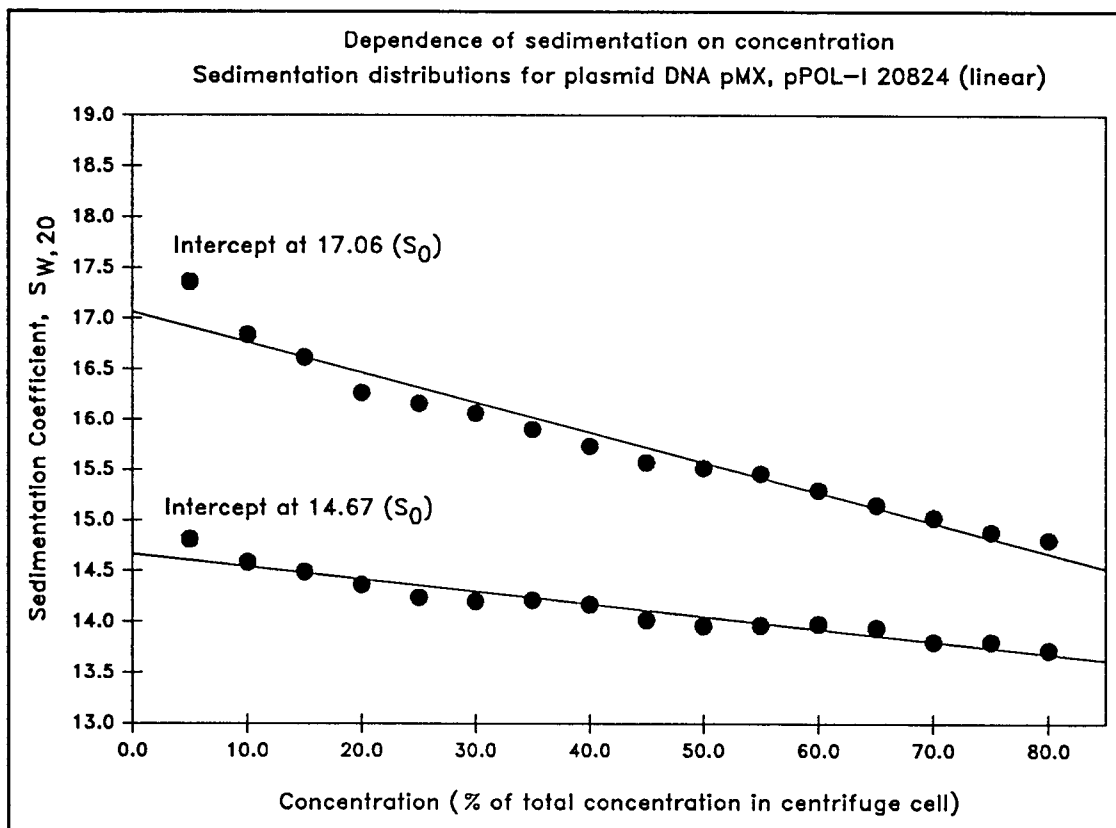
Figure 4.1.4: Extrapolation of sedimentation data to zero concentration. Shown are data for supercoiled plasmid pMX (S = 14.6) and linearized plasmid pPOL-1 208-24 (S = 17.06).

When plasmids were treated with ethidium bromide under low salt conditions, virtually no change in the sedimentation coefficient was observed. This is in contrast to studies performed in high salt (as high as 6.2 M Li^+ [26]), which show a clear change in sedimentation coefficient after various amounts of ethidium bromide have been added. Again, only the supercoiled (major component) averaged sedimentation coefficient is plotted, and the results are shown in Figure 4.1.5. It is possible that the ethidium bromide did not intercalate into the DNA, but rather bound to the DNA electrostatically because of the low salt concentration.
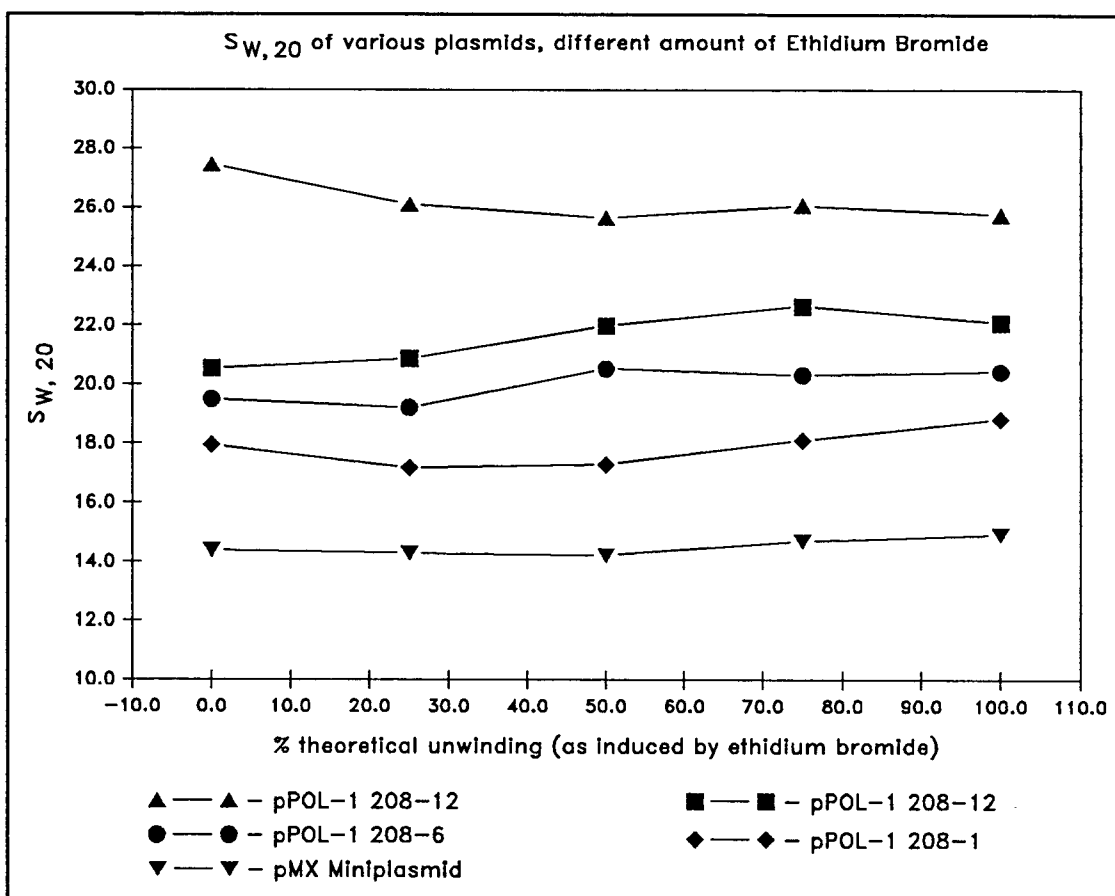
Figure 4.1.5: Effect of ethidium bromide on sedimentation velocity of several different plasmid DNA molecules in low salt.

4.2 Diffusion studies on plasmid DNA

To test the program LaserScan, plasmid DNA was measured by quasi-elastic light scattering. The analysis of the diffusion data was complicated by the presence of different conformations of the plasmid; a mixture exists which contains the linear, supercoiled and open circular conformation, also some concatamers might be present. Even small amounts of smaller DNA/RNA fragments left over from the plasmid purification seem to contribute to the autocorrelation spectrum.

It is possible to somewhat alleviate this complication by a double component fit. The diffusion coefficient derived from the second component fit agreed quite well with the diffusion

coefficient calculated from sedimentation velocity studies, as long as the samples were not too concentrated. Samples with higher concentrations displayed strong nonideality effects and it was impossible to obtain useful information.

Also, since it was not possible to determine accurately viscosity and refractive index of a solution, only data uncorrected for viscosity and refractive index are available. Two samples were analyzed in a diluted concentration to minimize effects caused by non-ideality, viscosity and refractive index changes. Calculated sedimentation coefficients for these samples matched closely the sedimentation coefficients experimentally determined by the van Holde - Weischet analysis. The results are shown in Table 4.2.

Sample	Molecular weight	$D_{20,w}$ $(* 10^{-8})$	experimental $S_{20,w}$ $(* 10^{-13})$	calculated $S_{20,w}$ $(* 10^{-13})$
pPOL-1 208-1 (diluted)	1884161.93	4.60	18.0	16.0
pPOL-1 208-6 (diluted)	2532765.73	3.82	19.5	17.9

Table 4.2: Calculated diffusion coefficient for diluted samples pPOL-1 208-1 and pPOL-1 208-6. The experimental measurement was performed by the van Holde - Weischet method.

4.3 Sedimentation studies on reconstituted pPOL-1 208-12

An experiment was designed to determine binding characteristics of nucleosomes to supercoiled DNA templates and if the addition of ethidium bromide to reconstituted plasmids would in any way affect the binding of nucleosomes to the templates through conformational changes induced into the chromatin structure by the ethidium bromide. Both conformational changes in the protein/DNA complex and molecular weight changes would affect the sedimentation pattern of the samples and could therefore be analyzed by analytical ultracentrifugation.

Initially, sedimentation experiments were performed on reconstituted pPOL-1 208-12. DNA was reconstituted with varying ratios (R) of nucleosomes per DNA molecule. An R value

of 1.0 is defined as one nucleosome per 208 bp DNA, the expected spacing pattern for a fully reconstituted 5S repeated DNA molecule.

Results indicated that nucleosomes would position more closely packed, because plasmids reconstituted with ratios higher than 1.0 did not exhibit aggregation. It has to be pointed out that the assumption of nucleosome spacing cannot be made alone based on ratios of histones/DNA calculated for the reconstitution. Inevitably, histones are lost during reconstitution by sticking to the walls of the dialysis tubing etc., therefore, the true histone/DNA ratio is probably lower.

Figure 4.3.1 shows van Holde - Weischet distributions for nucleosome reconstitutions of pPOL-1 208-12 with ratios between R = 0.3 and R = 1.5. Samples reconstituted with ratios of 1.2 and higher also show a small amount of larger aggregates (sedimenting with S-values of 75 and higher), ratios higher than R = 1.5 immediately precipitated out of solution because of aggregation. Free DNA exhibits a small amount of open circular or linear DNA.

It appears that intermediate ratios of nucleosomes/DNA produce a broad distribution of species, ranging between 30 and 55 S for an R value of 0.9, for example. Distributions become more narrow once the plasmid is increasingly loaded with nucleosomes. This observation is in contrast to the narrow distributions found in partially reconstituted linear DNA molecules [27]. It should be noted that an artificial sharpening of the distributions can result from concentration effects, which might cancel heterogeneity effects.
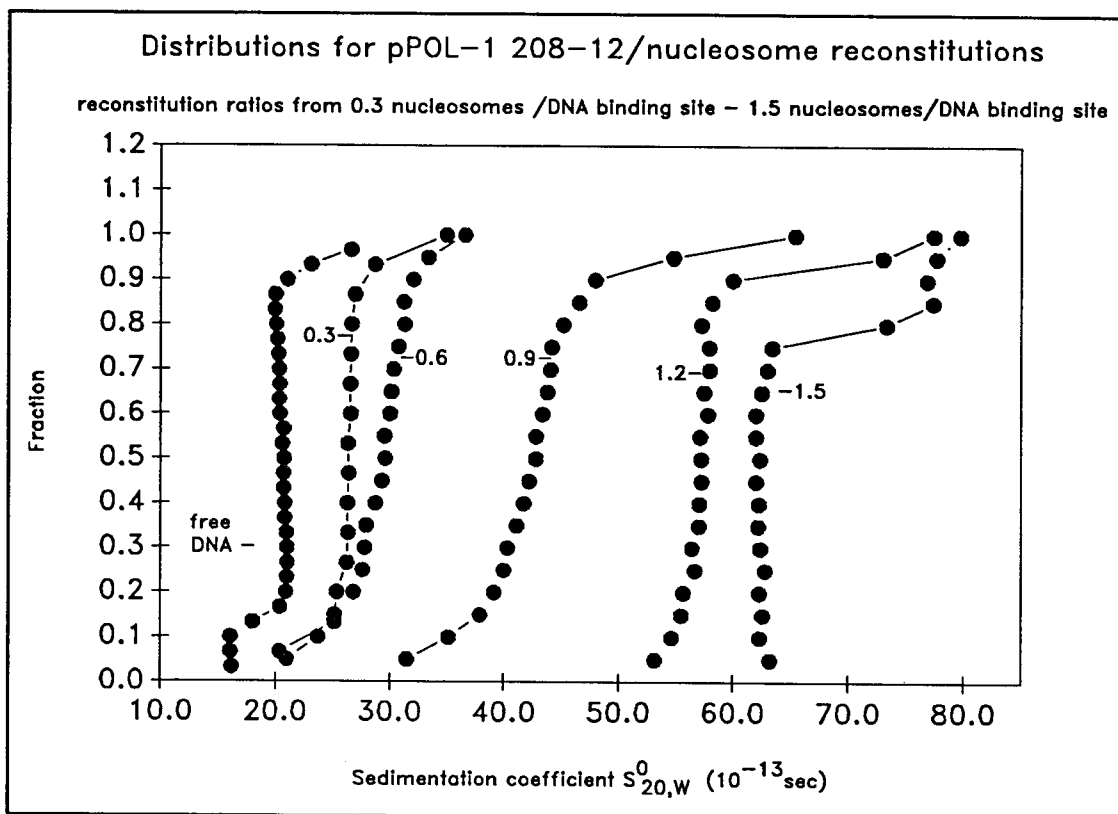
Figure 4.3.1: van Holde - Weischet distributions for sedimentation velocity experiments on pPOL-1 208-12 reconstituted with varying ratios of nucleosomes/DNA.

If ethidium bromide was added to reconstituted plasmids, two effects were observed:

(1) Distributions become more narrow for partially reconstituted plasmids

(2) Sedimentation velocity is reduced

The difference in distributions of reconstituted plasmid pPOL-1 208-12 with and without ethidium bromide is shown in Figures 4.3.2 - 4.3.5 for R values 0.3 to 1.2. The ethidium bromide concentration in the samples of Figures 4.3.2 - 4.3.5 was 0.117 mM, the DNA concentration was 45 μg/ml.
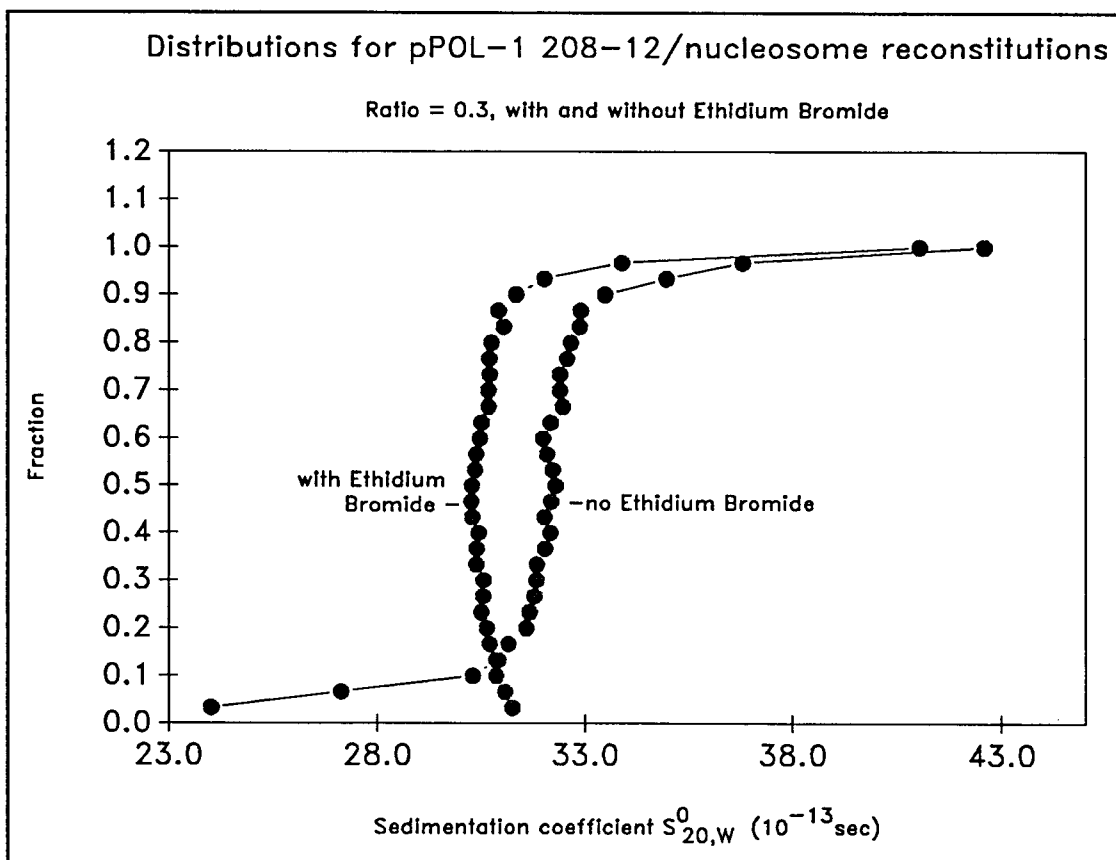
Figure 4.3.2: Effect of ethidium bromide on reconstituted plasmid DNA (pPOL-1 208-12 reconstituted with R = 0.3).
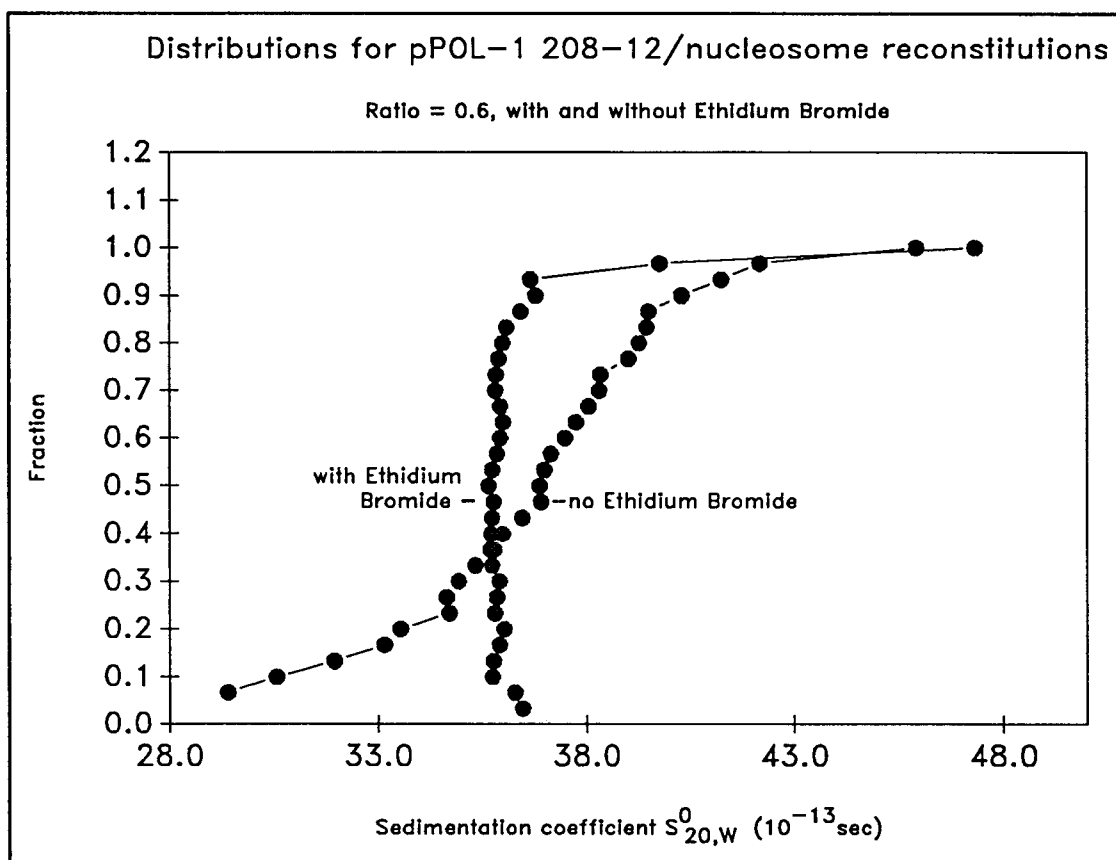
Figure 4.3.3: Effect of ethidium bromide on reconstituted plasmid DNA (pPOL-1 208-12 reconstituted with R = 0.6).
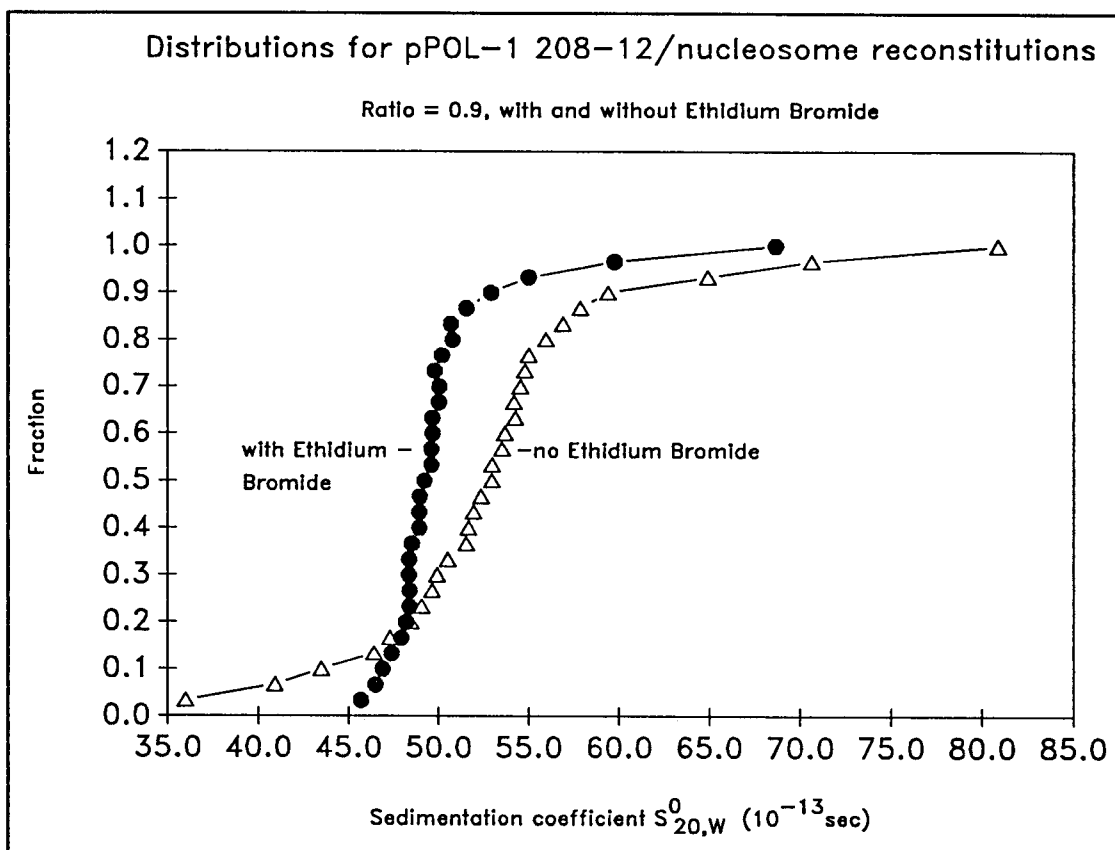
Figure 4.3.4: Effect of ethidium bromide on reconstituted plasmid DNA (pPOL-1 208-12 reconstituted with R = 0.9).

Reconstitutions of miniplasmid pMX allowed higher R values. The miniplasmid could be loaded with nucleosomes up to an R value of 2.0 before aggregation would occur. Distributions for the pMX plasmid reconstitutions at different R values showed the same characteristic behaviour as reconstitutions of pPOL-1 208-12: While free DNA and fully loaded DNA result in a narrow distribution, intermediate distributions are spread out over a large range, suggesting the presence of many different species. Again, this observation contrasts the narrow distributions found in linear DNA molecule reconstitutions [27]. The distributions are shown in Figure 4.3.6.

Since the pMX plasmid allowed R values of up to 2.0, nucleosomes probably bind to
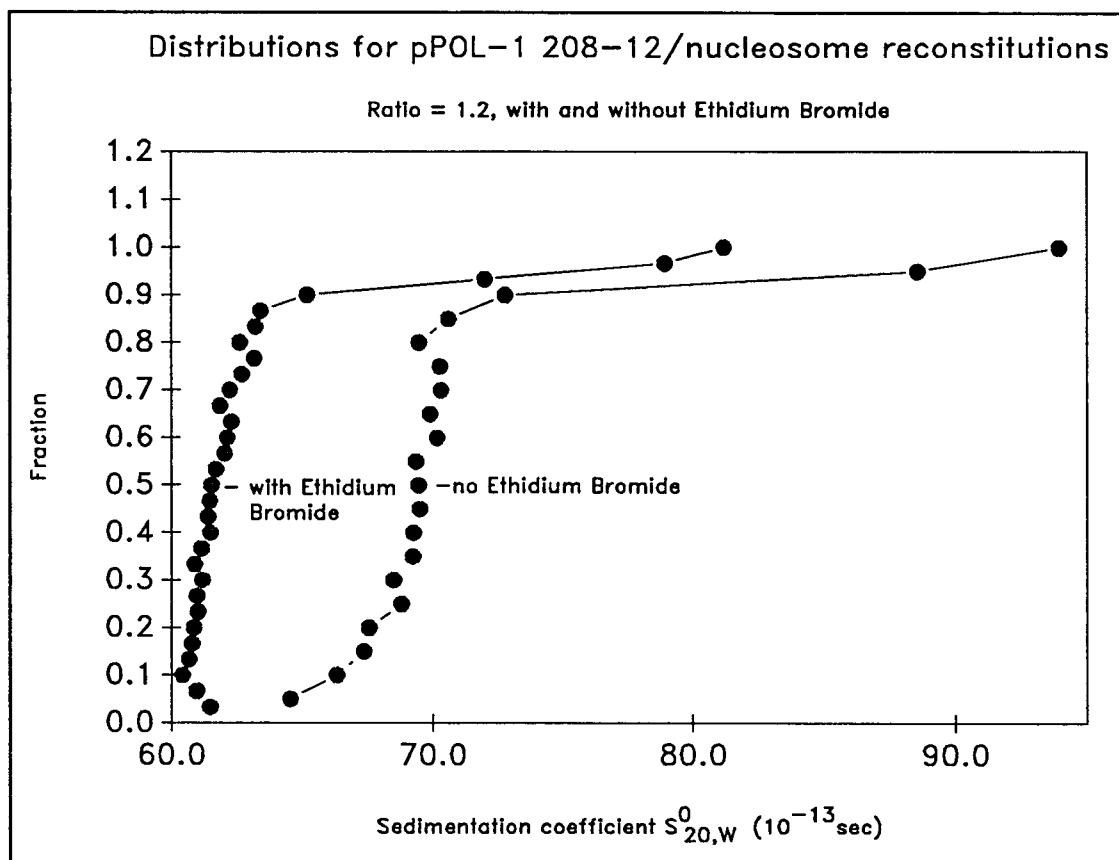
Figure 4.3.5: Effect of ethidium bromide on reconstituted plasmid DNA (pPOL-1 208-12 reconstituted with R = 1.2).

DNA stretches of around 130 bp (discounting lost histones during reconstitution). This correlates with tight dimers observed by [28]. The absence of known positioning sequences in the pMX DNA suggests a reason why a higher R value can be obtained for pMX than for pPOL-1 208-12: pPOL-1 208-12 contains 12 positioning sequences, which accounts for about half of the plasmid DNA. Since one half of the plasmid DNA can accommodate nucleosomes in a tight dimer packing, the packing of the other half could be determined by the 5S positioning signals suggesting an R value of about 1.5, which is observed.

When ethidium bromide was added to the reconstituted plasmid DNA, the same effect as in reconstituted pPOL-1 208-12 was observed:
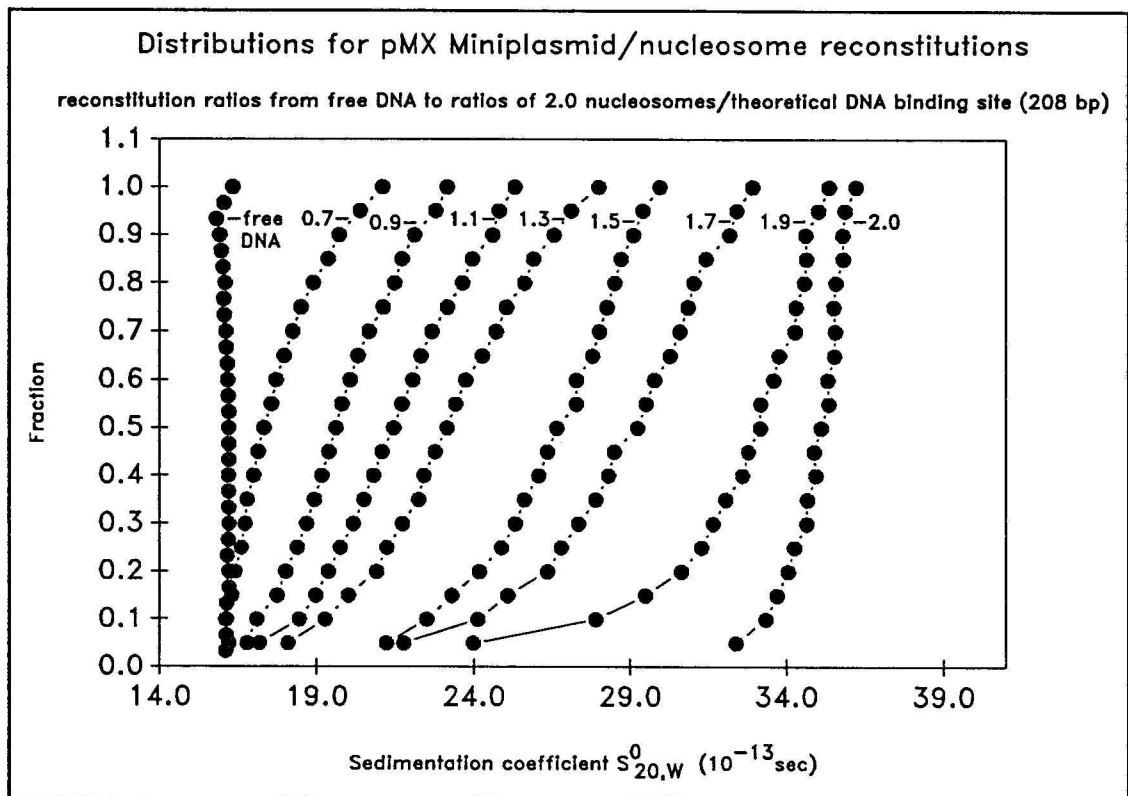
Figure 4.3.6: van Holde - Weischet distributions for reconstitution experiments of miniplasmid pMX DNA with ratios of 0.3 to 2.0.

(1) Distributions become more narrow for partially reconstituted plasmids

(2) Sedimentation velocity is reduced

These results are shown in Figure 4.3.7.
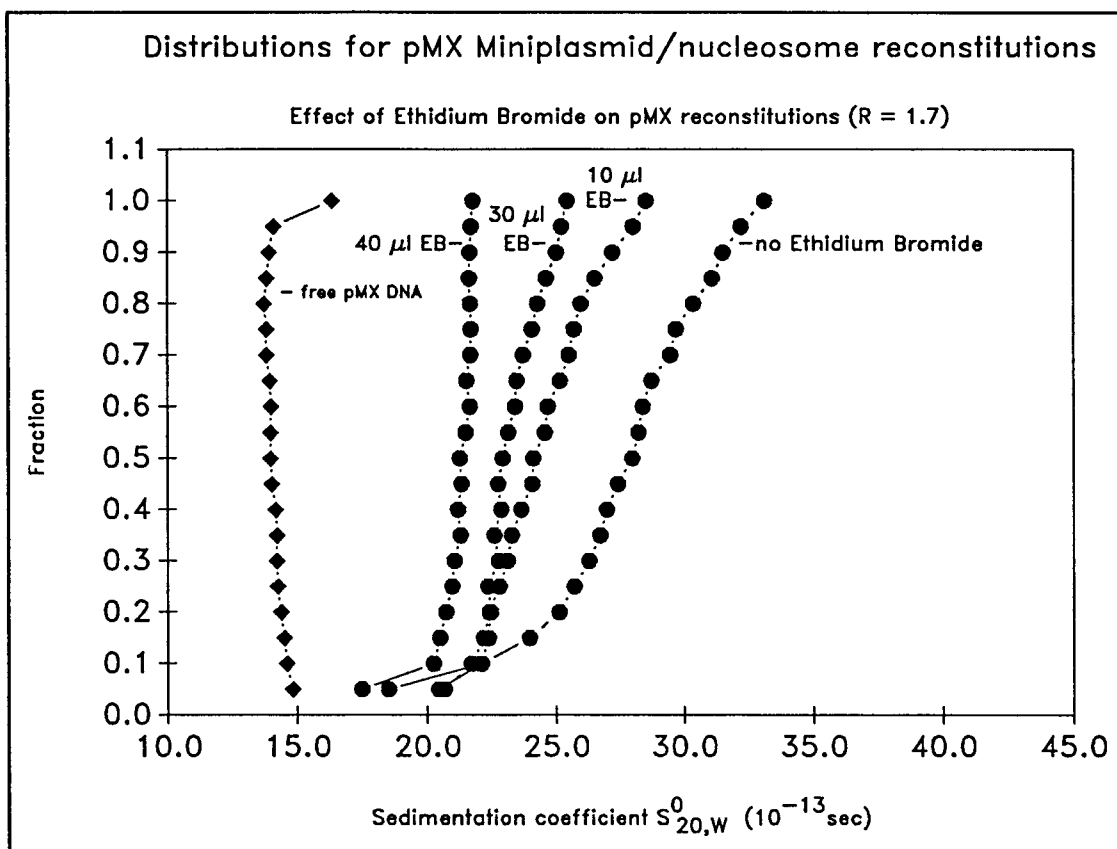
Figure 4.3.7: Effect of ethidium bromide on plasmid DNA reconstitutions (pMX reconstituted with R = 1.7)

The effect of ethidium bromide seems to be to either release nucleosomes because of unwinding and therefore to decrease the molecular weight or to open up the structure of the reconstituted particle to where the more extended conformation has a higher frictional coefficient.

Although free plasmid DNA did not exhibit any changes in sedimentation velocity upon addition of ethidium bromide (in low salt), a noticeable effect is observed for reconstituted plasmids. Nucleosomes, containing basic histones, can bind to DNA by neutralizing negative charges on the DNA backbone. This might enable ethidium bromide to intercalate, rather than to bind by electrostatic attraction. In low salt the DNA might not be completely neutralized by counterions, suggesting electrostatic binding of ethidium bromide to DNA.

The sharpening of the boundary observed upon addition of ethidium bromide is difficult to explain. van Holde - Weischet analysis implies a reduction of heterogeneity, which could result from a rapid exchange of histones from one template to another, until a stable distribution has been found which is more homogeneous than the sample without ethidium bromide. Further study of this phenomenon would be of interest.

5. MATERIALS AND METHODS

5.1 Materials

5.1.1 Chemicals

Unless otherwise noted, laboratory chemicals were purchased from Mallinckrodt, Sigma or Research Organics.

Acrylamide	Biorad
Bisacrylamide	Biorad
Ultrogel A2	LKB
SDS	BDH

5.1.2 Biochemicals

Restriction Enzymes	New England Biolabs, BRL, Pharmacia, Stratagene, IBI, Promega
T4 DNA Ligase	BRL
Mung Bean Nuclease	Pharmacia
Thermus Aquaticus DNA Polymerase	Promega
Ampicillin	Sigma
Chloramphenicol	Sigma
MacConkey Agar	Difco
Tryptic Soy Broth (TSB)	Difco
dNTP's	BRL
PCR Primers	Center for Gene Research, OSU
pUC sequencing primers	New England Biolabs
Geneclean	Bio 101

5.1.3 Buffers and solutions

TE	10 mM TrisHCl pH 8.0, 1 mM EDTA
1X TAE	40 mM Tris base, 40 mM acetic acid, 1 mM EDTA
1X TBE	89 mM Tris base, 89 mM boric acid, 1 mM EDTA
TSB-ampicillin	30g Tryptic Soy Broth, 1l ddH$_2$O, ampicillin (25 µg/ml), 1 ml 1% Thiamine
DNA Extraction solution 1	50 mM Glucose, 10 mM EDTA, 25 mM TRIS-HCL pH 8.0, 4 mg/ml Lysozyme
DNA Extraction solution 2	100 µl 10N NaOH, 500 µl 10% SDS, 4.4 ml ddH$_2$O (prepared fresh each time.)
DNA Extraction solution 3	5M NH$_4$OAC, 5M KOAC
Competent Cell Buffer	50 mM MgCl$_2$, 10 mM TrisHCl pH 8.0
Mung Bean nuclease reaction buffer	30 mM sodium acetate pH 5.0, 50 mM NaCl, 1mM ZnCl$_2$, 5% glycerol
1X Polymerase Chain Reaction Buffer (Stock buffer is 10X):	50 mM KCl, 10 mM Tris-HCl pH 8.4 (at 25°C), 2.5 mM MgCl$_2$, 100 µg/ml gelatin (Difco), 0.05% NP-40
DNAse free RNAse solution	500 µg RNAse A and 500 µg RNAse T in 1 ml TE pH 8.0, heated for 10 minutes in boiling water bath

5.1.4 Plasmids and strains

pUC 8 cloning vector	BRL

pAT153 [20]   our stock

XL1 blue   Stratagene

5.1.5 Others

Dialysis tubing   SPECTRA/POR, molecularporous

membrane tubing, M.W. cutoff:

12,000 - 14,000 (Spectrum Medical

Industries, Inc.)

5.2 Methods

5.2.1 Sterilization

Solutions, media, glass- and plasticware used for bacterial and DNA work were

sterilized at 125° for 30 minutes. Solutions not suitable for autoclaving were filtersterilized by

passing them through a 0.22 μ filter.

5.2.2 DNA precipitation

DNA was precipitated by adjusting the sample to 300 mM sodium acetate pH 5.2 and

adding 2 volumes of cold 95% ethanol. Short DNA fragments (<100 bp) and single stranded

DNA were precipitated by also adjusting the sample to 10 mM $MgCl_2$. The sample was mixed

by inversion and incubated at -20°C for 20 min, short pieces for 1 hour. Nucleic acids obtained

from bacterial extractions were preciptitated adding 1 volume isopropanol, mixing by inversion

and incubating for 1 hour at room temperature [28].

5.2.3 DNA purification

PCR fragments, restriction fragments and bacterial DNA were purified by a phenol and

chloroform/isoamyl alcohol extraction. Chloroform was mixed in a 24:1 (v/v) ratio with isoamyl

alcohol and stored at room temperature. Phenol was equilibrated with 1M Tris base solution to pH 8 and stored in a dark container at 4°C. For DNA purification, the sample was mixed with an equal volume of phenol, vortexed vigorously and centrifuged in a tabletop centrifuge until the phases were clearly separated. The aqueous phase was carefully removed and combined with an equal amount of chloroform/isoamyl alcohol mixture. The sample was vortexed and centrifuged to separate the aqueous from the organic phase. The aqueous phase was transferred to a clean tube and mixed with an equal amount of chloroform. After vortexing and phase separation, the sample was precipitated by ethanol precipitation [29].

5.2.4 Restriction digestion of DNA

All restriction digests were carried out in a buffer supplied by the manufacturer of the restriction enzyme. Usually, DNA solutions between 0.1 mg/ml - 1 mg/ml DNA were digested by incubating with the enzyme at 37°C for a minimum of 1 hour.

5.2.5 Removal of single stranded extensions

DNA with single stranded overhangs (3' or 5') was prepared for blunt-end ligation by treating with 3 units/µg DNA diluted Mung Bean nuclease in Mung Bean nuclease buffer. The reaction mixture was incubated for 1 hour at 30°C.

5.2.6 Purification of PCR primers

All PCR primers were synthesized on an Applied Biosystems DNA synthesizer at the Center for Gene Research and Biotechnology at OSU. Oligonucleotides were purified by electrophoresis on 20% acrylamide, 8 molar urea gels. The bands were located by UV shadowing and the top band, usually the darkest shadow, was excised from the gel. The gelslice was ground with a siliconized glassrod in a 50 ml steril plastic tube and 2 volumes of

TE pH8 were added to the gel powder. The suspension was incubated for 12 hours at 37°C with agitation. The gel particles were centrifuged to the bottom of the tube and the supernatant was carefully transferred to a clean tube. The extraction was repeated 2 times and supernatants were pooled. The DNA was recovered by ethanol precipitation [30].

5.2.7 Ligation of plasmid DNA and DNA fragments

In general, DNA ligations were performed by incubating 1 µg DNA with 5 units of T4 DNA ligase in 1X ligation buffer (supplied by the manufacturer) overnight at 20°C. Linearized plasmid DNA was circularized for transformation by incubating approximately 100-500 ng of plasmid DNA. Blunt-ended insert DNA was ligated to plasmid DNA by adding the insert DNA in a 10-30 molar excess over the plasmid DNA. Insert DNA with 5' or 3' overhangs (sticky ends) was performed by incubating insert DNA with plasmid DNA in a 3-5 molar excess. Ligation of polymerizing insert DNA was accomplished by incubating a 100 fold molar excess of insert DNA over plasmid DNA for 24 hours at 20°C.

5.2.8 Polymerase Chain Reaction (PCR)

PCR reactions were carried out basically as described in [31]. 1 µg of target DNA was mixed with 2 mM of each dATP, dCTP, dGTP and dTTP, 1 µM of each primer and 10 µl 10X PCR buffer. Steril ddH$_2$O was added to 99 µl reaction volume. The reaction mixture was heated to 98°C for 2 minutes to eliminate any DNA nuclease activity. 2.5 units (1 µl) of *Thermus Aquaticus* DNA polymerase was added, the reaction mix was overlaid with sterilized and purified mineral oil to avoid evaporation and 40 amplification cycles were performed:

Reaction cycles: Denaturation: 2   minutes at 94°C

                      Annealing: 2   minutes at 50°C

                      Elongation: 11 minutes at 71°C

Since the annealing temperature determines the stringency of the amplification (higher annealing temperatures will result in greater selectivity in the recovery of the targeted gene) the optimal annealing temperature is primer dependent and must be determined experimentally. After amplification, the mineral oil layer was removed with a steril transfer pipette and the DNA products were purified using chloroform extraction and ethanol precipitation. 10% of the reaction products were analyzed by 1% agarose gel electrophoresis on a minigel. Alternatively, reaction products were excised from the agarose gel slab and purified by the "Geneclean" method (adsorption of DNA molecules to glassbeads under high salt, elution with ddH$_2$O or low-salt buffer).

5.2.9 Preparation of competent cells

(Note: all laboratory ware, solutions, containers and chemicals for procedures involving bacteria were appropriately sterilized)

Competent cells were prepared by streaking 2-4 µl of a stock culture of XL-1 "blue" Escherichia coli (Stratagene) on a MacConkey Agar plate and incubating at 37°C until isolated colonies reached a diameter of 1-2 mm. A well isolated colony was picked with a loop and transferred to 5 ml 1X TSB and incubated with agitation overnight at 37°C. 100 µl of overnight grown culture were transferred to a 50 ml erlenmeyer flask containing 20 ml 1X TSB and incubated at 37°C with vigorous shaking until early log phase was reached.

1.5 ml culture fluid were dispensed into microfuge tubes and chilled on ice. The cell suspension was centrifuged at 5000 RPM at 4° C for 5 min. The supernatant was removed with a sterile transfer pipet and the tube was rechilled on ice. 500 µl cold 50 mM CaCl$_2$/10 mM TRIS-HCl pH 8 were added and the cell pellet was vortexed to homogeneity. The suspension was incubated for 30 min on ice and then centrifuged at 5000 RPM and 4° C for 5 min. The

supernatant was removed with a sterile transfer pipette and the pellet was resuspended with 0.2 ml cold 50 mM $CaCl_2$/10 mM TRIS-HCL pH 8, vortexed and stored at 4° C for 24 hours prior to transformation [29].

5.2.10 Transformation of vector DNA into _Escherichia coli_ competent cells

Prior to transformation, ligase in the ligation mix was inactivated by heating the ligation mixture at 65°C for 5 min. One aliquot of competent cells was combined with 5 ul of ligation mixture (ca. 500 ng DNA), vortexed, and incubated on ice for 45 minutes without any agitation. The transformation mixture was then carefully transferred to a 42°C waterbath and heatshocked for 120 seconds. Immediately after the heatshock treatment, the transformation mixture was incubated on ice for 90 seconds. 800 µl of prewarmed 2X TSB without any antibiotic was added and the mixture was incubated for 30 minutes at 37°C with shaking. 200 µl transformation mixture were plated on MacConkey agar plates containing 25 µg/ml ampicillin and incubated until colonies became visible [29].

5.2.11 Plasmid purification from _Escherichia coli_ cultures

Plasmid DNA preparations varied depending on the amount required. Small scale preparations (1.5 ml cultures) were prepared for colony screening. The typical plasmid DNA yield for a miniprep was 5-10 µg DNA, sufficient for several restriction digests and transformations. Large scale amplification was employed when large quantities of highly purified DNA was required for analytical experiments, such as analytical ultracentrifugation or quasi-elastic light scattering experiments. A large scale plasmid preparation typically yielded 10 mg of very pure plasmid DNA per liter of culture.

Small scale plasmid preparation:

The desired cell stock or colonies from transformation experiments were plated on MacConkey agar plates containing 25 μg/ml ampicillin. When colonies reached 1-2 mm in diameter, a well isolated colony was used to inoculate 5 ml TSB-ampicillin (25 μg/ml). The culture was incubated with vigorous shaking at 37°C overnight. Cells were harvested by centrifugation in a Beckman J-6B preparative centrifuge. After decanting the supernatant and blotting off excess culture media by standing centrifuge tubes upside down on absorbent paper, cells were lysed by resuspending the cell pellet in 200 μl DNA extraction solution I and incubating for 15 minutes at room temperature. Then, 400 μl of DNA extraction solution II were added to the lysate and the solution was vortexed vigorously.

After incubating for 10 min on ice, 200 μl of icecold DNA extraction solution III were added and the mixture was vortexed vigorously to break up cells. The solution was incubated on ice for another 20 minutes, during which the solution was repeatedly vortexed. After incubation, the solution was centrifuged for 30 minutes at 4,200 rpm and 4°C in a Beckman J-6B preparative centrifuge.

The supernatant was carefully transferred to a clean 1.5 ml eppendorf tube and 1 volume of isopropanol were added. The solution was mixed by inversion and incubated for 1 hour at room temperature. The solution was centrifuged for 10 minutes at 14,000 rpm in a tabletop centrifuge. After slightly drying the pellet, nucleic acids were reconstituted in 100 μl TE pH 8.0 containing 1% RNAse A and T. The solution was incubated at room temperature for 1 hour. After incubation, the solution was ethanol precipitated [29].

Large scale plasmid preparation:

The desired cell stock was plated on MacConkey agar plates containing 25 μg/ml ampicillin. When colonies reached 1-2mm in diameter, a well isolated colony was used to inoculate 5 ml TSB-ampicillin (25 μg/ml). The culture was incubated with vigorous shaking

at 37°C overnight. 100 µl of overnight grown culture were then transferred to a 50 ml erlenmeyer culture flask containing 20 ml TSB-ampicillin and incubated for 6 hours at 37°C with vigorous shaking. Optical density measurements at 340 nm to assay the status of culture growth were never performed.

The entire contents were added to a 5 liter culture flask containing 1 l TSB-ampicillin. After 6 hours of incubation at 37°C with vigorous shaking, 500 µl chloramphenicol stock (34 mg/ml 70% ethanol) were added to the culture and the culture was incubated for an additional 12 hours. The cells were harvested by centrifuging 1 l culture for 20 minutes at 4,200 rpm and 4°C in a Beckman J-6B preparative centrifuge. After decanting the supernatant and blotting off excess culture media by standing centrifuge bottles upside down on absorbant paper, cells were lysed by resuspending the cell pellet in 14 ml DNA extraction solution I and incubating for 15 minutes at room temperature. Then, 28 ml of DNA extraction solution II were added to the lysate and the solution was vortexed vigorously.

After incubating for 10 min on ice, 14 ml of icecold DNA extraction solution III were added and the mixture was vortexed vigorously to break up cells. The solution was incubated on ice for another 20 minutes, during which the solution was repeatedly vortexed. After incubation, the solution was centrifuged for 30 minutes at 4,200 rpm and 4°C in a Beckman J-6B preparative centrifuge.

The supernatant was carefully transferred to a clean 250 ml centrifuge bottle and 1 volume of isopropanol was added. The solution was mixed by inversion and incubated for 1 hour at room temperature. The solution was centrifuged for 20 minutes at 8,000 rpm and 4°C in a Sorvall preparative ultracentrifuge. After decanting the supernatant, the pellet was washed with icecold 70% ethanol and re-centrifuged. After slightly drying the pellet, nucleic acids were reconstituted in 8 ml TE pH 8.0 and 50 µl RNAse solution was added and the solution was

incubated at room temperature for 1 hour. After incubation, the solution was transferred to a 30 ml Corex tube and ethanol precipitated. The precipitate was washed with icecold 70% ethanol and recentrifuged.

At this point, the plasmid DNA was further purified by CsCl density equilibrium centrifugation. The cell pellet was resuspended in 8.2 ml TE pH 8.0 and 8.8 g CsCl (optical grade) and 0.82 ml 20 mg/ml ethidium bromide were added, mixed and the density of the solution was checked. It should be around 1.54 g/ml. Two 5.1 ml Quickseal tubes (Beckman) were filled with the solution, loaded into a VTi 65.2 rotor and spun at 45,000 rpm for 17 hours. The band corresponding to the supercoiled plasmid was located by shining 300 nm UV light on the polyallumer tube and removed from the tube with a syringe by puncturing the tube wall with a needle underneath the band. Care was taken not to transfer any other band from the tube.

The plasmid solution from both tubes was pooled and the ethidium bromide removed by repeated extractions with isoamyl alcohol, until no trace of pink color remained in the supernatant. CsCl was removed from the sample by dialyzing for 12 hours against 2 liter TE pH 8.0 at 4°C. A_{260} absorption measurements were taken to determine the concentration of the DNA. The plasmid purity was checked by loading a small sample on a 1% agarose gel. If necessary, another CsCl density equilibrium centrifugation was performed to eliminate traces of RNA or linear or nicked plasmid DNA [29].

5.2.12 Nucleosome reconstitution on DNA templates

H1 and H5 free histone octamers were prepared from chicken erythrocytes by the method of Hansen et al. [32]. A 2 M NaCl solution was prepared containing plasmid DNA template (50 µg/ml). The solution was combined with histone octamers in different molar ratios. Throughout, the solution was kept cold on ice. One molar ratio "R" is defined as one histone octamer per 200 bp DNA template. Thus, for a 1000 bp DNA molecule, R = 1 corresponds to

5 histone octamers per DNA molecule. A program (HISDNA) was written to calculate the necessary volumes of DNA, histone and NaCl solutions at given concentrations at different values of R. A copy of the source code is appended to this document.

After mixing, the solution was transferred to molecularporous membrane tubing and dialyzed by continuous dilution against TE pH 8.0 from 2.0 M NaCl 0.0 M NaCl. The dialysis was carried out at 4°C and performed as follows: The dialysis tubing was placed in a 500 ml flask containing 500 ml TE pH 8.0, 2.0 molar in NaCl. A peristaltic pump was used to pump a solution of TE pH 8.0 without NaCl into the first flask at a constant flow rate of 1.5 ml/minute. An overflow in the first flask assured constant volume and high speed stirring provided uniform salt concentration. At the end of a 24 hour period, the dialysis bag was transferred to a 500 ml flask containing TE pH 8.0 and the solution was dialyzed for an additional 6 hours. The dilution profile can be modeled by following differential equation:

$$(17) \quad \frac{dX}{dt} = -\frac{f}{V} * X, \quad \text{with the solution:} \quad X(t) = C * \exp\left(-\frac{f}{V} * t\right)$$

where X is the NaCl concentration at time t, f is the flowrate in ml/min, C is the initial NaCl concentration and V is the volume of the dialysis flask. The profile for 2 different flow rates is shown below:

Dialysis dilution profile
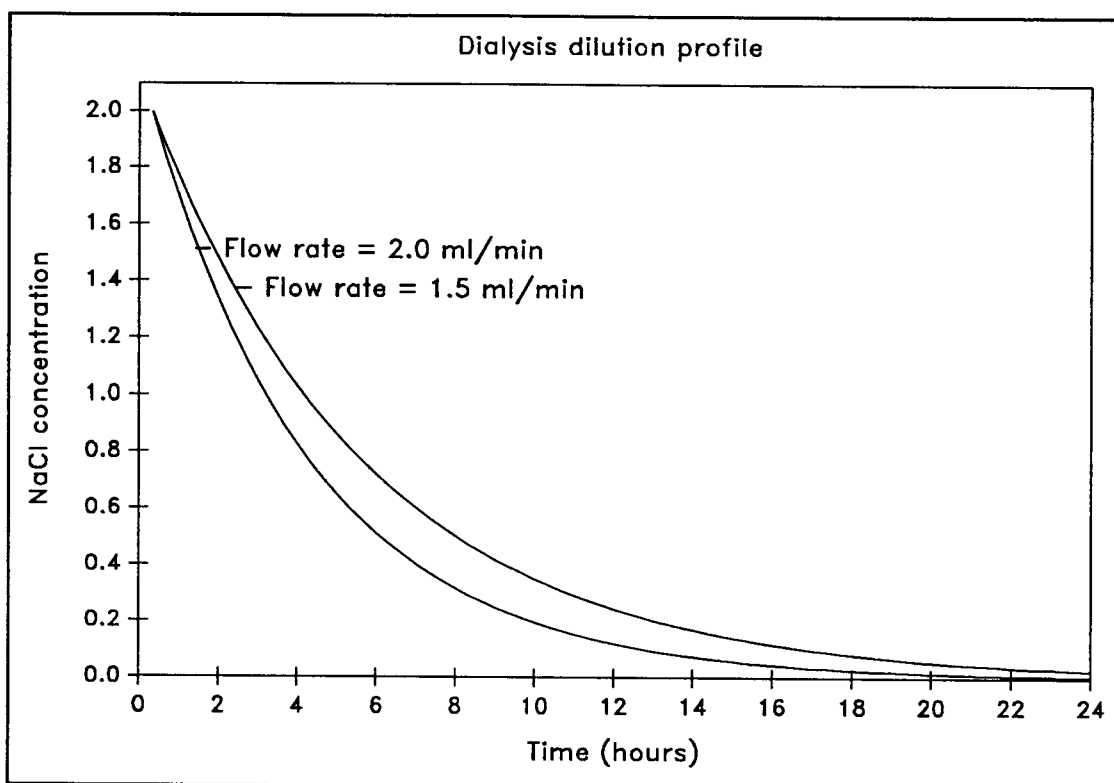
Figure 5.1: Dilution profile for salt dialysis of nucleosome reconstitutions. Two different flow rates are shown.

This method was chosen for its reproducibility and reliability. It was compared to the step dialysis method of Hansen et al., 1990 by dialyzing a nucleosome reconstitution of 12-mer repeat of a 5S sea urchin gene. The reconstitution was checked by sedimentation velocity analysis and the S - value distributions were shown to be in agreement.

5.2.13 "Geneclean" method

The "Geneclean" method [33] was used to purify DNA fragments from agarose gels. The desired band was carefully excised from the gel with a sterilized razorblade and transferred to a 1.5 ml microfuge tube. 1.5-2 volumes of saturated NaI were added to the tube and the tube was incubated at 55°C for 5 minutes of until the gelslice was melted. 10 µl of glassmilk (Bio

101) for each microgram of DNA were added to the solution and the suspension was thoroughly vortexed.

After incubation on ice for 10 minutes the glassmilk was pelleted by centrifugation in a tabletop centrifuge for 10 seconds. The supernatant was discarded and the glassmilk was resuspended in 450 μl "New Wash" (Bio 101) by vortexing and pelleted again by centrifugation. The washing procedure was repeated 3 times and the final centrifugation was repeated once to remove all residual washing solution from the glassmilk pellet. The DNA was extracted from the pellet by resuspending the pellet in 10 μl of ddH$_2$O or 10 μl low-salt TE buffer, incubating at 55°C for 5 minutes, pelleting the glassmilk by centrifugation and saving the supernatant. The extraction was repeated twice and the supernatants were pooled.

6. BIBLIOGRAPHY

(1). van Holde, K.E. (1985) *Physical Biochemistry* Prentice Hall, Inc. Englewood Cliffs, NJ 07632, 103-136.

(2). ibid, p 111.

(3). C.R. Cantor and P.R. Schimmel (1980) *Biophysical Chemistry Part II: Techniques for the Study of Biological Structure and Function* W.H Freeman and Company, San Francisco, 623-627.

(4). ibid, p 603.

(5). van Holde, K. E. & Weischet, W. O. (1978) *Biopolymers* **17**, 1387-1403.

(6). C.R. Cantor and P.R. Schimmel (1980) *Biophysical Chemistry Part II: Techniques for the Study of Biological Structure and Function* W.H Freeman and Company, San Francisco, p 618.

(7). K.E. van Holde (1985) *Physical Biochemistry* Prentice Hall, Inc. Englewood Cliffs, NJ 07632, pp 114, 125.

(8). ibid, p 125.

(9). Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986) *Numerical Recipes - the Art of Scientific Computing* Cambridge University Press, Cambridge pp 523-528.

(10). ibid, 625-626.

(11). Robinson, N. C., & Talbert, L. (1986) *Biochemistry*, **25**, 2328-2334.

(12). K.E. van Holde (1985) *Physical Biochemistry* Prentice Hall, Inc. Englewood Cliffs, NJ 07632, p 117.

(13). ibid, p 114.

(14). Portions of this program have been obtained from: Bevington, P. R. (1969) *Data Reduction and Error Analysis for the Physical Sciences* McGraw-Hill Book Company, New York, pp 235-239.

(15). Portions of this routine have been obtained from: Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986) *Numerical Recipes - the Art of Scientific Computing* Cambridge University Press, Cambridge pp 523-528. pp 277-282.

(16). ibid, 106-107.

(17). Jakeman, E., & Pike, E. R. (1969), *J. Phys. A: Math., Nucl., Gen.* **2**, 411

(18). Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986) *"Numerical Recipes - the Art of Scientific Computing"* Cambridge University Press, Cambridge pp 529-538.

(19). Messing, J. (1983) *Methods in Enzymology*, **101**, 20-89.

(20). Simpson, R. T., Thoma, F. & Brubaker, J. M. (1985) *Cell* **42**, 799-808

(21). Paule, M. R., Iida, C. T., Perna, P. J., Harris, G. H., Brown-Shimer, S. L., & Kownin, P. (1984) *Biochemistry* **23**, 4167-4172.

(22). Kownin, P., Iida, C. T, Brown-Shimer, S., & Paule, M. R. (1985) *Nuc. Acids Res.* **13**, 6234-6247.

(23). Kownin, P., Bateman, E., & Paule, M. R. (1987) *Cell* **50**, 693-699.

(24). NIH Genbank, entry for pBR 322 plasmid DNA.

(25). Bethesda Research Laboratories (1985) *"Catalogue & Reference Guide"*, Gaithersburg, Maryland.

(26). Baase, W. A., & Johnson, W. C. (1979) *Nucl. Acids Res.* **6-2**, 797-814.

(27). Hansen, J., University of Texas, Health Science Center, San Antonio, unpublished results.

(28). Tatchell, K., & van Holde, K. E. (1978) *Proc. Natl. Acad. Sci. USA* **75-8**, 3583-3587.

(29). Maniatis, T., Fritsch, E. F., & Sambrook, J. (1983) *"Molecular Cloning. A Laboratory Manual"*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

(30). Giovanonni, S. Oregon State University, personal communication.

(31). Saiki, R. K., Gelford, D. H., Stoffel, S. (1988) *Science* **239**, 487-494.

(32). Hansen, J.C., Ausio, J., Stanik, V. H., & van Holde, K. E. (1989) *Biochemistry* **28**, 9129-9136.

(33). BIO 101, BOX 2284, La Jolla, California 92038

7. APPENDICES

APPENDIX A - PROGRAM ULTRASCAN

TABLE OF CONTENTS:

PROGRAM DESCRIPTION:

General:

This program is intended for use with the Beckman Analytical Ultracentrifuge Model-E. The interface is designed and constructed by Carl Baker, Baker Industry Labs, Corvallis, Oregon. The analog to digital conversion is performed by the CROM-1 A/D conversion board from Metrabyte Corporation. The program is written in several independent modules to minimize memory requirements. The program allows concurrent usage of the Beckman stripchart recorder during data collection. The program allows chaining among modules and can be interrupted by the keystroke combination <CTRL><BREAK>. The user interface is menu-driven and menus can be aborted using the <ESC> key.

Hardware requirements:

The program is intended to run on an IBM PC/AT style computer with a 80-*86 type processor and preferably equipped with a math co-processor (80-287/387). It requires 640 KB RAM and can make use of a RAM drive. A line printer is optional. The program is suitable for graphics screendumps after loading a graphics screen driver (i.e. GRAPHICS.COM). A harddisk of at least 10 MB is required. The program itself requires 2 MB of harddisk space. The program can run under all MS-DOS versions 2.* or later. A mouse is required for editing data and for some of the data analysis routines.

Installation:

To install the program, place the master disk into the floppy drive and type:

B:\>install <ENTER>

The program will then extract all archived files and install itself on the harddisk in a subdirectory called "C:\US17>". Please read the chapter on PROGRAM CONFIGURATION to determine how to set up your program.

The configured program is started by typing:

C:\US17>US <ENTER>

It is advisable to load a graphics screen driver before starting the program. A suitable screen driver is graphics.com for EGA adaptor boards. A commercially available program such as PIZAZZ Plus will provide greater flexibility and is recommended for usage as graphics screen handler.

If you have a floating-point math coprocessor of the type 80287 or 80387 installed in your computer, you can take advantage of the increased speed provided by the coprocessor. The file "BLINEC.EXE" contains a special instruction set for calculations in the equilibrium data analysis routine, optimized for the 80287/80387 math coprocessor. The analysis program will only recognize the file named "BLINE.EXE", therefore, if you want to take advantage of the increased speed of calculation, you have to rename the file "BLINEC.EXE" to "BLINE.EXE".

File Format

The file format is as follows:

<run identification, scan number,., edit status, cell number>

where:
1) run identification is simply the name of your run
   (6 characters)
2) scan number is the number of your scan (1-20)
3) edit status can be "R" for raw data or "E" for edited data
4) cellnumber is the number of the cell being scanned.

3) and 4) are in the extension of the filename, separated by a period from the name and scan number. All files are binary record files, except configuration files with file extensions ".DAT", ".TYP" and ".CFG". Archived files with the extension ".ARC" are binary encoded as well. Other files contain information about configuration (CROM1.CFG and CROM2.CFG and files with extension ".INF"). Rotor information is stored in files ROTOR.DAT and ROTOR.TYP.

Help files are ASCII formatted and have extensions ".HLP". Files with extension ".XCL" are exclusion files. They contain information about scans excluded during editing from a set of acquisitions. For export of data files into different programs via ASCII format, you have to use the CONVERT PROGRAM. Look at the help topic "Convert Program" for more information.

Metafiles can be generated from the van Holde - Weischet and equilibrium analysis programs. The format for the metafiles is ASCII (extension ".MET"), with the data columns comma delimited for easy import into programs like SIGMAPLOT, QUATTRO PRO, SUPERCALC etc. For detailed information on the contents of the metafiles, please consult the respective ANALYSIS PROGRAM chapter.

Results from the analysis methods are written to files with extensions ".RES". The filename consists of the run number and the cell number: <run number, cell number,".RES">.

ULTRASCAN PROGRAM

The ULTRASCAN program is the main program containing the main menu:

```
┌─────────────────────────────────────────────────┐
│        Ultracentrifuge Data Analysis Program:    │
├─────────────────────────────────────────────────┤
│                                                   │
│         ┌─────────────────────────┐               │
│         │       MAIN MENU:        │               │
│         ├─────────────────────────┤               │
│         │ F1: Data Acquisition    │               │
│         │ F2: View + Edit Data    │               │
│         │ F3: Analysis Methods    │               │
│         │ F4: Select Data         │               │
│         │ F5: Utility Programs    │               │
│         │ F6: Help Program        │               │
│         │ F7: Exit Program        │               │
│         └─────────────────────────┘               │
│                                                   │
└─────────────────────────────────────────────────┘
```
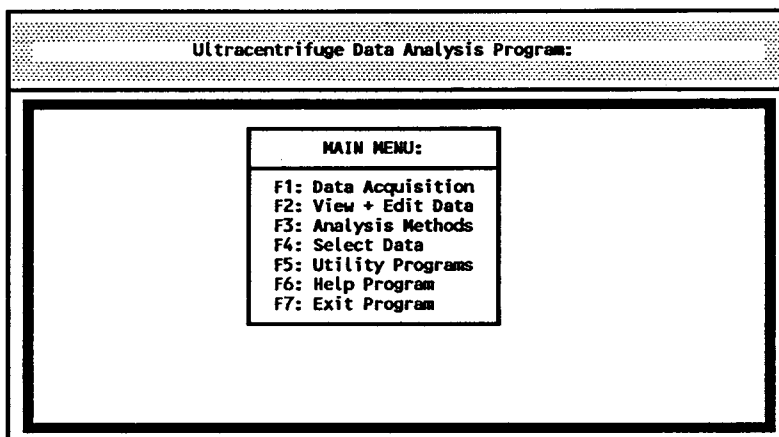
This menu allows the user to select between the main sub-programs of the ULTRASCAN program:

1. Data Acquisition
2. Viewing and editing data previously acquired
3. Analysis of the edited data
4. Data identification and selection of subdirectories for storage of data
5. Utility program subroutines
6. Help program for on-line help documentation for the various functions and programs
7. Exiting the ULTRASCAN program.

The sub-programs are selected by using the cursor keys and hitting <ENTER>. Alternatively, the indicated function keys can be pressed as a shortcut.

HELP PROGRAM:

The help program has information and help on all functions available in the program:

```
┌─────────────────────────────────────────────────────┐
│           Model-E Help Documentation:               │
│                                                     │
│  ┌────────────────────────────────────────────────┐ │
│  │                                                │ │
│  │        ┌──────────────────────────┐            │ │
│  │        │      HELP TOPICS:         │           │ │
│  │        ├──────────────────────────┤            │ │
│  │        │ F1: Data Acquisition      │           │ │
│  │        │ F2: View + Edit Data      │           │ │
│  │        │ F3: Analysis Programs     │           │ │
│  │        │ F4: Utility Programs      │           │ │
│  │        │ F5: CROM Switch Info      │           │ │
│  │        │ F6: Model-E Operation     │           │ │
│  │        │ F7: Exit Help             │           │ │
│  │        └──────────────────────────┘            │ │
│  │                                                │ │
│  └────────────────────────────────────────────────┘ │
└─────────────────────────────────────────────────────┘
```
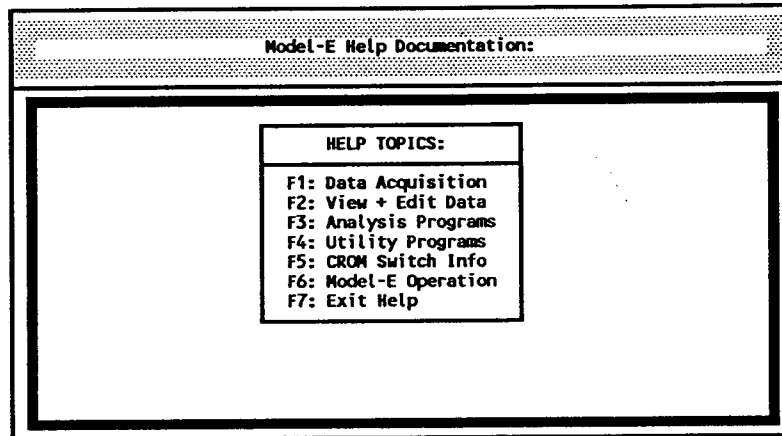
Different help topics are available for the analysis programs:

```
┌─────────────────────────────────────────────────────┐
│           Model-E Help Documentation:               │
│                                                     │
│  ┌────────────────────────────────────────────────┐ │
│  │                                                │ │
│  │       ┌──────────────────────────┐             │ │
│  │       │   ANALYSIS HELP TOPICS:    │            │ │
│  │       ├──────────────────────────┤             │ │
│  │       │ F1: Weischet Analysis      │           │ │
│  │       │ F2: 2nd Moment Analysis    │           │ │
│  │       │ F4: Equilibrium Analysis   │           │ │
│  │       │ F5: Exit Analysis Help     │           │ │
│  │       └──────────────────────────┘             │ │
│  │                                                │ │
│  │                                                │ │
│  └────────────────────────────────────────────────┘ │
└─────────────────────────────────────────────────────┘
```

and the utility programs:

```
┌─────────────────────────────────────────────────────────────┐
│                 Model-E Help Documentation:                  │
│                                                              │
├──────────────────────────────────────────────────────────────┤
│  ┌─────────────────────────────┐                            │
│  │ UTILITY HELP TOPICS:        │                            │
│  ├─────────────────────────────┤                            │
│  │ F1: Desktop Utility         │                            │
│  │ F2: Archiving Program       │                            │
│  │ F3: Configuration           │                            │
│  │ F4: Data Selection          │                            │
│  │ F5: File Conversion         │                            │
│  │ F6: File Format             │                            │
│  │ F7: Rotor Data Entry        │                            │
│  │ F8: Exit Utility Help       │                            │
│  └─────────────────────────────┘                            │
│                                                              │
└──────────────────────────────────────────────────────────────┘
```
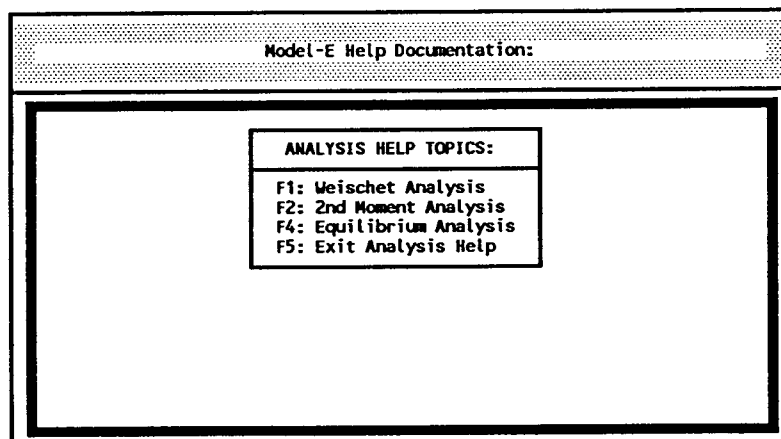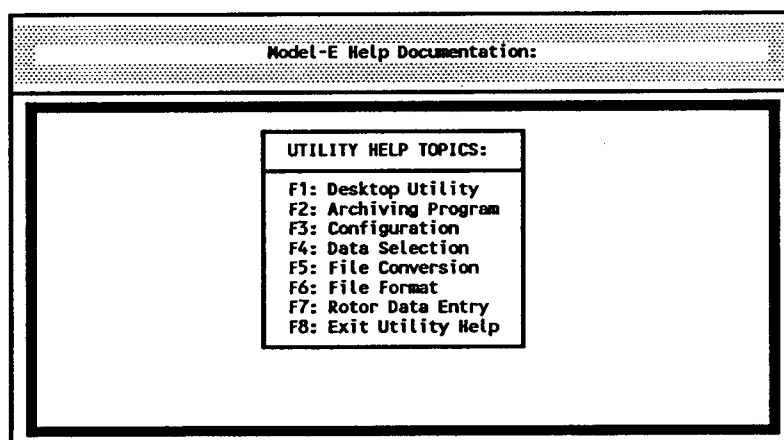
DATA ACQUISITION PROGRAM:

The data acquisition program allows the user to interface directly with the Model-E and collect the data from the photo-multiplier tube. The stripchart recorder can be used concurrently. This program is intended to accumulate data from the ultracentrifuge over the interface and analog to digital conversion board in your computer.

```
┌─────────────────────────────────────────────────────────┐
│        Ultracentrifuge Data Acquisition:                │
├─────────────────────────────────────────────────────────┤
│                                                         │
│              ┌──────────────────────┐                   │
│              │  Parameter setting:  │                   │
│              ├──────────────────────┤                   │
│              │  F1: Automatic Mode  │                   │
│              │  F2: Manual Mode     │                   │
│              │  DEFAULT: Automatic  │                   │
│              │  F4: Utilities       │                   │
│              │  F5: Exit            │                   │
│              └──────────────────────┘                   │
│                                                         │
└─────────────────────────────────────────────────────────┘
```
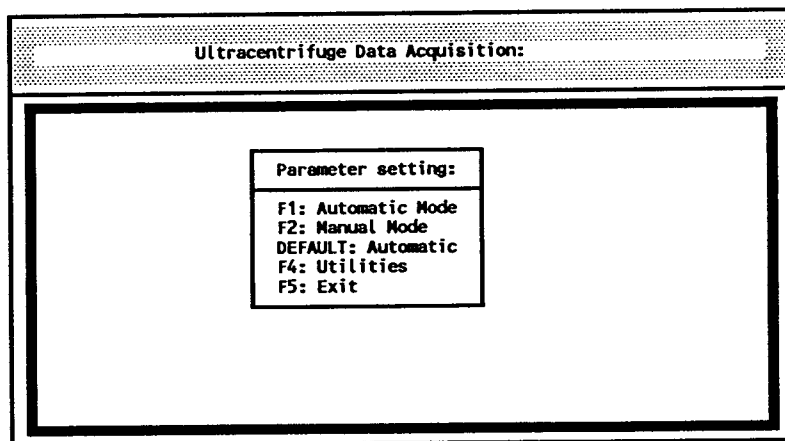
You have the option of a manual and an automatic data acquisition mode. Choose automatic mode (default) for normal operation and manual mode for test runs. Also available from this menu are the utility programs for program configuration and setup of acquisition parameters. The program lets you define a path and a run identification name, where your files will be stored. If you choose the manual mode, the program will prompt you for a valid filename (the datapath and the name of the file where you want your data to be stored. A valid filename consists out of a maximum of 8 characters, a period and a 3 character extension, no spaces are allowed). Note: if you make more than one run with the same filename, the old file will be overwritten. The default is the automatic mode: The program will list the configuration from the last run and prompts you for any changes.

Configuration items are:

1) run identification - this allows you to give your data a common name which will be used to identify every scanned data with this particular experiment. 6 characters maximum are allowed.

2) number of cells - how many cells you are scanning.

3) number of scans - how many scans you are performing on each cell.

4) data path - this is the drive and name of the subdirectory, where your data will be stored. This way you can separate your data from that of a co-worker.

5) Cell description - enter a unique description for the contents of each cell. The data is written to a file <run name.INF>.

After establishing this information, it will be written to a file (CROM2.CFG) and you can start the data acquisition. You also have the option to abort the run by pressing the escape key <ESC>. Then the program calls a driver program (AD-DATA.EXE), which communicates with the ultracentrifuge. Basically, it waits for a trigger from the ultracentrifuge and starts accumulating data into a filename with following format:

where run identification is simply the name of your run (6 characters), scan number is the number of your scan (1-20), edit status can be "R" for raw data or "E" for edited data and cellnumber is the number of the cell being scanned. There is no need to supervise the program while it is running, no action is required, until the run is finished. The program will display the status of the acquisition at all times by letting you know which cell and which scan is being scanned and how many more will be scanned. After the run, several additional information need to be entered:

1) Rotor speed in revolutions per minute

2) RTIC value

3) Rotor used for the run.

The program will look up temperature calibration data in a rotor information table (rotor.dat) and combined with your RTIC value calculate the true temperature. After running the data acquisition, run the edit program.

DATA SELECTION:

DATA SELECT HELP MANUAL:

With this function you can select the data and the subdirectory search path of the current run identification. This information can be changed in the beginning of each program. Once selected, it stays current throughout each program, or until changed. The data is saved in an ASCII file with the name of the run identification and extension ".DAT". It contains information about the number of cells, number of scans, the rotor used, and the search path of the storage for the data.

```
┌─────────────────────────────────────────────────────────────┐
│              Parameter Configuration:                        │
├─────────────────────────────────────────────────────────────┤
│  ┌─────────────────────────────────────────────────────────┐ │
│  │ Current Configuration:                                  │ │
│  │                                                         │ │
│  │ Run ID (6 Char.) = TEST01                               │ │
│  │ Number of cells  =  3                                   │ │
│  │ Number of scans  =  5                                   │ │
│  │ File datapath     = D:\DATA\                            │ │
│  │ Cell information:                                       │ │
│  │                                                         │ │
│  │ Cell #1: Sample #1                                      │ │
│  │ Cell #2: Sample #2                                      │ │
│  │ Cell #3: Sample #3                                      │ │
│  │                                                         │ │
│  │                                                         │ │
│  │ Do you want to change this configuration? (y/n)         │ │
│  │                                                         │ │
│  └─────────────────────────────────────────────────────────┘ │
└─────────────────────────────────────────────────────────────┘
```

The configuration from the last run will be shown, by entering "y", this configuration can be adapted to the new run. Following screen will appear:

```
┌─────────────────────────────────────────────────────────────┐
│              Parameter Configuration:                        │
├─────────────────────────────────────────────────────────────┤
│  ┌─────────────────────────────────────────────────────────┐ │
│  │ Current Configuration:                                  │ │
│  │                                                         │ │
│  │ Run ID (6 Char.) = TEST01    --- <--┌──────────────┐    │ │
│  │ Number of cells  =  3        ---    │ Change ? (y/n)│    │ │
│  │ Number of scans  =  5        ---    └──────────────┘    │ │
│  │ File datapath     =          ---                        │ │
│  │ D:\DATA\                                                │ │
│  │ Cell information:            ---                        │ │
│  │                                                         │ │
│  │ Cell #1: Sample #1                                      │ │
│  │ Cell #2: Sample #2                                      │ │
│  │ Cell #3: Sample #3                                      │ │
│  │                                                         │ │
│  └─────────────────────────────────────────────────────────┘ │
└─────────────────────────────────────────────────────────────┘
```

Enter "y" to change the configuration at the desired point. To change the run identification, enter "y" at this point and following screen will appear:

```
┌─────────────────────────────────────────────────────────┐
│                 Parameter Configuration:                │
├─────────────────────────────────────────────────────────┤
│                                                         │
│  Current Configuration:                                 │
│                                                         │
│  Run ID (6 Char.) = TEST01        ---                   │
│  Number of cells  =  3            ---                   │
│  Number of scans  =  7            ---                   │
│  File datapath     =              ---                   │
│  D:\DATA\                                               │
│                                                         │
│                                                         │
│                                                         │
│  Please enter a new run-ID. You have 6 characters to enter a unique │
│  identification for your run. The run-ID will be the first 6 characters │
│  of the filename of the file containing the ASCII data from your run. │
│                                                         │
│  New Run Identification: TEST02                         │
│                                                         │
└─────────────────────────────────────────────────────────┘
```

Enter a new run identification at the prompt. You have 6 characters maximum for the run identification.

```
┌─────────────────────────────────────────────────────────┐
│                 Parameter Configuration:                │
├─────────────────────────────────────────────────────────┤
│                                                         │
│  Current Configuration:                                 │
│                                                         │
│  Run ID (6 Char.) = TEST02        ---                   │
│  Number of cells  =  3            ---                   │
│  Number of scans  =  5            ---                   │
│  File datapath     =              ---                   │
│  D:\DATA\                                               │
│  Cell information:                ---                   │
│                                                         │
│                                                         │
│  Please enter a new run-ID. You have 6 characters to enter a unique │
│  identification for your run. The run-ID will be the first 6 characters │
│  of the filename of the file containing the ASCII data from your run. │
│                                                         │
│  Is this a new run? (y/n)                               │
│                                                         │
└─────────────────────────────────────────────────────────┘
```

If there is no previous data with the same identification, the program will assume the identification for the run to be for a new run. If the selected data has been archived and not extracted, a warning will be displayed:

```
┌─────────────────────────────────────────────────────────────┐
│                   Parameter Configuration:                    │
├─────────────────────────────────────────────────────────────┤
│                                                               │
│  Current Configuration:                                       │
│                                                               │
│  Run ID (6 Char.) = TEST02          ---                       │
│  Number of cells  =  3              ---                       │
│  Number of scans  =  5              ---                       │
│  File datapath    =                 ---                       │
│  D:\DATA\                                                     │
│  Cell information:                  ---                       │
│                                                               │
│                                                               │
│  Please enter a new run-ID. You have 6 characters to enter a unique │
│  identification for your run. The run-ID will be the first 6 characters │
│  of the filename of the file containing the ASCII data from your run. │
│  WARNING! You have to de-archive this data first or change the directory │
│  Hit any key to continue...                                   │
│                                                               │
└─────────────────────────────────────────────────────────────┘
```

Next, the number of cells and scans should be entered:

```
╔══════════════════════════════════════════════════════════════════╗
║                    Parameter Configuration:                        ║
╠══════════════════════════════════════════════════════════════════╣
║                                                                    ║
║  Current Configuration:                                            ║
║                                                                    ║
║  Run ID (6 Char.) = TEST02          ---    ┌──────────────────┐   ║
║  Number of cells  =  3              ---    │ Number of cells: │   ║
║  Number of scans  =  5              ---    │                  │   ║
║  File datapath     =                ---    │ F1:  1 cell      │   ║
║  D:\DATA\                                  │ F2:  2 cells     │   ║
║  Cell information:                  ---    │ F3:  3 cells     │   ║
║                                            │ F4:  4 cells     │   ║
║                                            │ F5:  5 cells     │   ║
║  Cell #1:                                  │ F6:  6 cells     │   ║
║  Cell #2:                                  └──────────────────┘   ║
║  Cell #3:                                                          ║
║                                                                    ║
║                                                                    ║
╚══════════════════════════════════════════════════════════════════╝
```

```
╔══════════════════════════════════════════════════════════════════╗
║                    Parameter Configuration:                        ║
╠══════════════════════════════════════════════════════════════════╣
║                                                                    ║
║  Current Configuration:                                            ║
║                                                                    ║
║  Run ID (6 Char.) = TEST02          ---                           ║
║  Number of cells  =  3              ---                           ║
║  Number of scans  =  5              ---    How many scans? (max=20) 10
║  File datapath     =                ---                           ║
║  D:\DATA\                                                          ║
║  Cell information:                  ---                           ║
║                                                                    ║
║  Cell #1:                                                          ║
║  Cell #2:                                                          ║
║  Cell #3:                                                          ║
║                                                                    ║
║                                                                    ║
╚══════════════════════════════════════════════════════════════════╝
```

A-15

The datapath can be entered to specify a new location for the disk storage of the collected data. This way you can separate your data from the data of your co-worker:

```
╔═════════════════════════════════════════════════════════════╗
║                    Parameter Configuration:                  ║
╚═════════════════════════════════════════════════════════════╝
┌─────────────────────────────────────────────────────────────┐
│  Current Configuration:                                      │
│                                                              │
│  Run ID (6 Char.) = TEST01          ---                      │
│  Number of cells  =  3              ---                      │
│  Number of scans  =  5              ---                      │
│  File datapath    =                 ---                      │
│  D:\DATA\                                                    │
│  Cell information:                  ---                      │
│                                                              │
│                                                              │
│  Please enter a new default directory path where your        │
│  datafiles are stored. Enter a valid path!                   │
│                                                              │
│  New path: D:\TEST                                           │
└─────────────────────────────────────────────────────────────┘
```

If an invalid datapath is entered, an error message is displayed:

```
╔═════════════════════════════════════════════════════════════╗
║                    Parameter Configuration:                  ║
╚═════════════════════════════════════════════════════════════╝
┌─────────────────────────────────────────────────────────────┐
│  Current Co┌──────────────────────────────────────┐          │
│            │     You have to choose a valid path! │          │
│  Run ID (6 └──────────────────────────────────────┘          │
│  Number of cells  =  3              ---                      │
│  Number of scans  =  5              ---                      │
│  File datapath    =                 ---                      │
│  D:\DATA\                                                    │
│  Cell information:                  ---                      │
│                                                              │
│                                                              │
│  Please enter a new default directory path where your        │
│  datafiles are stored. Enter a valid path!                   │
│                                                              │
│  New path: D:\TEST                                           │
└─────────────────────────────────────────────────────────────┘
```

In addition to the run identification, a sample description for the contents of each cell can be included. This information is helpful to identify your run, if the data analysis is preformed at a later time. You will be prompted for a description for each cell:

```
╔══════════════════════════════════════════════════════════════════╗
║                       Parameter Configuration:                     ║
╠══════════════════════════════════════════════════════════════════╣
║  ┌──────────────────────────────────────────────────────────────┐ ║
║  │ Current Configuration:                                        │ ║
║  │                                                               │ ║
║  │ Run ID (6 Char.) = TEST02          ---                        │ ║
║  │ Number of cells  = 3               ---                        │ ║
║  │ Number of scans  = 5               ---                        │ ║
║  │ File datapath    =                 ---                        │ ║
║  │ D:\DATA\                                                      │ ║
║  │ Cell information:                  ---                        │ ║
║  │                                                               │ ║
║  │ Cell #1:                                                      │ ║
║  │ Cell #2:                                                      │ ║
║  │ Cell #3:                                                      │ ║
║  │                                                               │ ║
║  │ Please enter a new description for cell #1:                   │ ║
║  │ ?                                                             │ ║
║  └──────────────────────────────────────────────────────────────┘ ║
╚══════════════════════════════════════════════════════════════════╝
```

When all information is properly entered, the new configuration can be written to disk. The information is stored in two different files, "RUN-ID.DAT" and "CROM2.DAT". If you want to modify the setup once more, answer "n" to the next prompt:

```
╔══════════════════════════════════════════════════════════════════╗
║                       Parameter Configuration:                     ║
╠══════════════════════════════════════════════════════════════════╣
║  ┌──────────────────────────────────────────────────────────────┐ ║
║  │ Current Configuration:                                        │ ║
║  │                                        Exit? (y/n)            │ ║
║  │ Run ID (6 Char.) = TEST02          ---                        │ ║
║  │ Number of cells  = 3               ---                        │ ║
║  │ Number of scans  = 10              ---                        │ ║
║  │ File datapath    =                 ---                        │ ║
║  │ D:\DATA\                                                      │ ║
║  │ Cell information:                  ---                        │ ║
║  │                                                               │ ║
║  │ Cell #1: Sample #1                                           │ ║
║  │ Cell #2: Sample #2                                           │ ║
║  │ Cell #3: Sample #3                                           │ ║
║  │                                                               │ ║
║  │                                                               │ ║
║  └──────────────────────────────────────────────────────────────┘ ║
╚══════════════════════════════════════════════════════════════════╝
```

CELL AND SCAN SELECTION:

For editing, viewing and data analysis it is sometimes useful to look at a portion of the scans and cells only. Therefore, ranges for the cells and scans to be manipulated can be determined with the cell/scan select function:

```
Please select the cells and scans you want to include in the analysis:

Current selection:

First cell:  1
Last cell :  3
First scan:  1
Last scan :  7

Change? (y/n)
```

Simply enter "y" to change the default setting, which is always set to the maximum number of cells and scans. Failure to provide more than one scan (which is necessary for data analysis) will result in an error message.

```
Please sel    You have to select more than one scan!    analysis:

Current selection:

First cell:  1
Last cell :  3
First scan:  4
Last scan :  4

Change? (y/n)
```

UTILITY PROGRAM:

The utility program has various functions to allow greater control and provides special features:

```
┌─────────────────────────────────────────────────────┐
│         Ultracentrifuge Data Analysis Program:        │
├─────────────────────────────────────────────────────┤
│  ┌─────────────────────────────────────────────────┐ │
│  │                                                   │ │
│  │          ┌──────────────────────────┐            │ │
│  │          │      UTILITY MENU:        │            │ │
│  │          ├──────────────────────────┤            │ │
│  │          │  F1: Desktop Utility      │            │ │
│  │          │  F2: Archiving Utility    │            │ │
│  │          │  F3: Program Configuration│            │ │
│  │          │  F4: Select Data          │            │ │
│  │          │  F5: Convert File Format  │            │ │
│  │          │  F6: Enter Rotor Data     │            │ │
│  │          │  F7: Escape to DOS Shell  │            │ │
│  │          │  F8: Help                 │            │ │
│  │          │  F9: Exit Utility Program │            │ │
│  │          └──────────────────────────┘            │ │
│  │                                                   │ │
│  └─────────────────────────────────────────────────┘ │
└─────────────────────────────────────────────────────┘
```
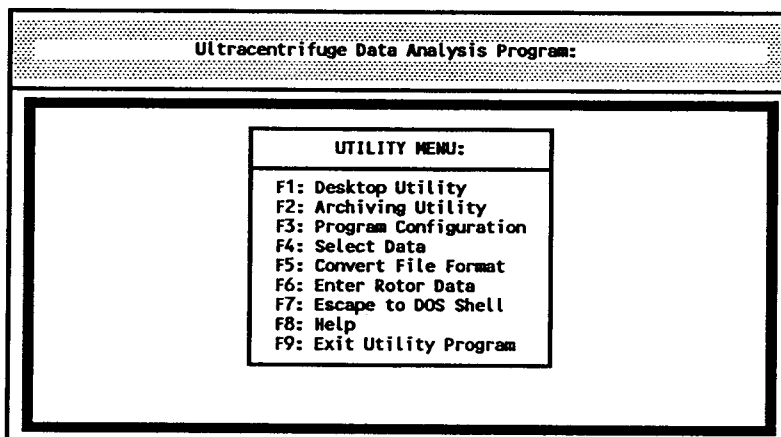
From this menu the various functions and features can be selected. Each function is described in detail in its own chapter.

DESKTOP UTILITY:

This program is intended to help you organize the data on your disk. It should allow functions like editing, copying, deleting and other file management options. It is not supplied with this program, but it can be added with the configuration program utility, so chaining to the program is possible. To include this option, you should add the full path and filename (including extension) of the desktop utility you want to use, when prompted by the configuration program.

ARCHIVE UTILITY:

The archiving program is intended to archive datafiles into a more compact format. This will help to conserve disk space. Archives can be extracted at any time when the information is needed again. The archive utility will archive your data into an archive with the name of your run identification, with an ".ARC" file extension. To archive a set of datafiles, simply enter the name of your run identification and the directory name and path, where the data is stored, using "Select Data". Select "Create Archive" to compress the data and "Extract Archive" to extract a previously archived set of datafiles. Use "View Archive" to inform yourself of the contents of an archive.

```
┌──────────────────────────────────────────────────────┐
│               Data Archiving Program:                 │
├──────────────────────────────────────────────────────┤
│  ┌────────────────────────────────────────────────┐  │
│  │        ┌───────────────────────────┐           │  │
│  │        │     ARCHIVING MENU:        │           │  │
│  │        │  F1: Select Data           │           │  │
│  │        │  F2: Create Archive        │           │  │
│  │        │  F3: Extract Archive       │           │  │
│  │        │  F4: View Archive          │           │  │
│  │        │  F5: Desktop Utility       │           │  │
│  │        │  F6: Help                  │           │  │
│  │        │  F7: Exit Program          │           │  │
│  │        └───────────────────────────┘           │  │
│  │                                                │  │
│  └────────────────────────────────────────────────┘  │
└──────────────────────────────────────────────────────┘
```

The archiving program will prompt you for the necessary information. To create an archive, enter the run identification, the name of the drive and path where the source data is located, then enter drive and path for the destination of the archive.

```
┌──────────────────────────────────────────────────────┐
│                Run Number Selection:                  │
├──────────────────────────────────────────────────────┤
│  ┌────────────────────────────────────────────────┐  │
│  │  Please select the Run Identification: TEST01   │  │
│  │                                                │  │
│  │                                                │  │
│  │                                                │  │
│  │                                                │  │
│  │                                                │  │
│  │                                                │  │
│  │                                                │  │
│  │                                                │  │
│  │                                                │  │
│  └────────────────────────────────────────────────┘  │
└──────────────────────────────────────────────────────┘
```

A-21

```
┌─────────────────────────────────────────────────────────────────┐
│░░░░░░░░░░░░░░░░░░░░░░░░░Run Number Selection:░░░░░░░░░░░░░░░░░░░░░░│
│░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░│
├─────────────────────────────────────────────────────────────────┤
│                                                                   │
│   Please select the Source Drive where the Data is located:       │
│                                                                   │
│      ┌──────────────────────┐                                     │
│      │  SELECT DRIVE:       │                                     │
│      ├──────────────────────┤                                     │
│      │  F1: Floppy Drive A: │                                     │
│      │  F2: Floppy Drive B: │                                     │
│      │  F3: Hard Disk       │                                     │
│      └──────────────────────┘                                     │
│                                                                   │
│                                                                   │
│                                                                   │
└─────────────────────────────────────────────────────────────────┘
```

If you have the data stored on the fixed drive, enter the subdirectory where the data is located:

```
┌─────────────────────────────────────────────────────────────────┐
│░░░░░░░░░░░░░░░░░░░░░░░░░Run Number Selection:░░░░░░░░░░░░░░░░░░░░░░│
│░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░│
├─────────────────────────────────────────────────────────────────┤
│                                                                   │
│   This is the current source directory: D:\DATA\                  │
│                                                                   │
│   Do you want to change it (y/n)? y                               │
│                                                                   │
│                                                                   │
│                                                                   │
│                                                                   │
│   Please enter a new default directory path where your            │
│   datafiles are stored. Enter a valid path!                       │
│                                                                   │
│   New path: D:\TEST                                               │
│                                                                   │
└─────────────────────────────────────────────────────────────────┘
```

A-22

REMEMBER:

Disk space is not saved, unless you delete the files of the run. The archive utility will not delete any data for safety reasons. You can delete data with your favorite desktop utility. For more info on this topic, view the Desktop help information.

```
                          Archiving Utility:


   Your files D:\DATA\5049*.* have been stored in archive:

   D:\DATA\5049.ARC

   If you want to conserve disk space, delete these files now.
   Press any key to continue...
```
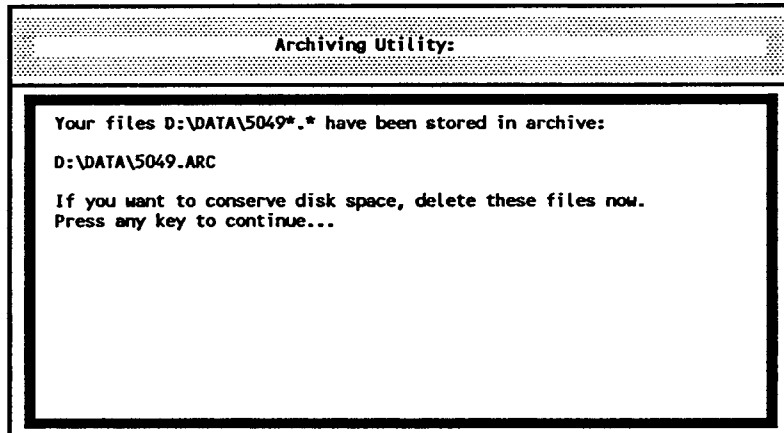
PROGRAM CONFIGURATION:

In order to run the program properly, various parameters need to be adjusted to gain optimal performance. This is accomplished in the configuration program. Following parameters need to be set:

```
:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
                    UCF Data Analysis Program Configuration:
:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::

Current Configuration:
Readings/second  = 50   ------ <--  Change ? (y/n)
Run Duration     = 2    min --
Input Channel    = CH0  ------
Maximum Range    = 10v  ------
Trigger Code     = 8    ------
Relay 0 Start    = on   ------
Relay 0 End      = off  ------
Relay 1 Start    = off  ------
Relay 1 End      = off  ------
I/O Address      = &H300 ----
Interrupt Level  = 5    ------
Graphics Adapter = EGAHIRES -
Desktop Utility  = ----------
c:\pctools\pcshell.exe
```

1) Readings per second (data acquisition rate)

2) Run duration (length in minutes of each scan)

3) Input channel from your analog to digital board

4) Voltage range of input (maximum voltage allowed in input)

5) Trigger code (binary code for trigger interrupt on board)

6) - 9) Relay activation settings (not required at present configuration)

10) I/O address (base address of your computer for input/output operations)

11) Interrupt level (which interrupt your computer uses)

12) Graphics adapter (which graphics adapter and monitor is connected to your computer)

13) Desktop utility path (drive and subdirectory information for your favorite desktop program)

A-24

Readings/sec and Run duration:

```
.............................................................................
              UCF Data Analysis Program Configuration:
.............................................................................

  Current Configuration:
  Readings/second  = 50  ------
  Run Duration     = 2   min --     ┌─────────────────────┐
  Input Channel    = CHO ------     │  Readings/second:   │
  Maximum Range    = 10v ------     ├─────────────────────┤
  Trigger Code     = 8   ------     │  F1:  0.1/sec       │
  Relay 0 Start    = on  ------     │  F2:  0.5/sec       │
  Relay 0 End      = off ------     │  F3:  1/sec         │
  Relay 1 Start    = off ------     │  F4:  2/sec         │
  Relay 1 End      = off ------     │  F5:  5/sec         │
  I/O Address      = &H300 ----     │  F6:  10/sec        │
  Interrupt Level  = 5   ------     │  F7:  20/sec        │
  Graphics Adapter = EGAHIRES -     │  F8:  50/sec        │
  Desktop Utility  = ----------     │  F9:  100/sec       │
  c:\pctools\pcshell.exe            └─────────────────────┘
```

```
.............................................................................
              UCF Data Analysis Program Configuration:
.............................................................................

  Current Configuration:
  Readings/second  = 10  ------    New Run Length in Minutes:1.5
  Run Duration     = 2   min --
  Input Channel    = CHO ------
  Maximum Range    = 10v ------
  Trigger Code     = 8   ------
  Relay 0 Start    = on  ------
  Relay 0 End      = off ------
  Relay 1 Start    = off ------
  Relay 1 End      = off ------
  I/O Address      = &H300 ----
  Interrupt Level  = 5   ------
  Graphics Adapter = EGAHIRES -
  Desktop Utility  = ----------
  c:\pctools\pcshell.exe
```

The product of readings/sec and run duration should not exceed 10.000 to avoid memory problems with your computer. 10.000 is the number of maximum data pairs/scan allowable in the program. You can set readings/sec to a high value, like 50 or 100, if you want very good resolution. The run duration on an average 286 AT computer will actually take approximately twice as long for 50 readings/sec or 4 times as long for 100 readings/sec as specified in run duration, because the computer has to process the data in addition to accumulating it.

Attention:

Therefore, you have to be careful that the run duration does not exceed the maximum time set by the multiplexer from the ultracentrifuge for the interval between scans (the program has to be ready and waiting for the next trigger, BEFORE the multiplexer is started to give the trigger for the next run or you will loose some or all data and the file order for accumulation will be mixed up.

A-25

Channel:

```
┌──────────────────────────────────────────────────────────────┐
│ ┌──────────────────────────────────────────────────────────┐ │
│ │              UCF Data Analysis Program Configuration:     │ │
│ └──────────────────────────────────────────────────────────┘ │
│                                                                │
│  ┌──────────────────────────────────────────────────────────┐ │
│  │ Current Configuration:                                    │ │
│  │ Readings/second  = 10   ------                            │ │
│  │ Run Duration     = 1.5 min --      ┌─────────────────────┐ │
│  │ Input Channel    = CH0 ------      │ Channel Selection:  │ │
│  │ Maximum Range    = 10v ------      │                     │ │
│  │ Trigger Code     = 8    ------     │ F1: Channel 0       │ │
│  │ Relay 0 Start    = on   ------     │ F2: Channel 1       │ │
│  │ Relay 0 End      = off ------      │                     │ │
│  │ Relay 1 Start    = off ------      │                     │ │
│  │ Relay 1 End      = off ------      │                     │ │
│  │ I/O Address      = &H300 ----      │                     │ │
│  │ Interrupt Level  = 5    ------     │                     │ │
│  │ Graphics Adapter = EGAHIRES -      │ DEFAULT = CH0       │ │
│  │ Desktop Utility  = ----------      └─────────────────────┘ │
│  │ c:\pctools\pcshell.exe                                    │ │
│  └──────────────────────────────────────────────────────────┘ │
└──────────────────────────────────────────────────────────────┘
```

Your CROM-1 board has two channels where it can accumulate data on (channel 0 and channel 1). Therefore, you can use another instrument (e.g: an HPLC scanner) with your CROM-1 board. The program needs to know from which channel your CROM-1 board is supposed to accumulate data. The interface box is built to interface with channel 0, therefore, keep this set to channel 0, unless you want to change it to another system.

Maximum range:

```
┌─────────────────────────────────────────────────────────────┐
│              UCF Data Analysis Program Configuration:         │
│                                                               │
├─────────────────────────────────────────────────────────────┤
│  Current Configuration:                                       │
│  Readings/second  = 10   ------                               │
│  Run Duration     = 1.5 min --     ┌───────────────────────┐  │
│  Input Channel    = CH0 ------     │   Maximum Range:      │  │
│  Maximum Range    = 10v ------     │                       │  │
│  Trigger Code     = 8   ------     │  F1:  1 Volt          │  │
│  Relay 0 Start    = on  ------     │  F2:  2 Volt          │  │
│  Relay 0 End      = off ------     │  F3:  5 Volt          │  │
│  Relay 1 Start    = off ------     │  F4: 10 Volt          │  │
│  Relay 1 End      = off ------     │                       │  │
│  I/O Address      = &H300 ----     │                       │  │
│  Interrupt Level  = 5   ------     │                       │  │
│  Graphics Adapter = EGAHIRES -     │  DEFAULT = 10 Volt    │  │
│  Desktop Utility  = ----------     └───────────────────────┘  │
│  c:\pctools\pcshell.exe                                       │
└─────────────────────────────────────────────────────────────┘
```

The interface system has been designed to provide 10 V DC current for data acquisition. Unless you use the board with another interface, this setting should be left at 10 V.

Trigger code:

```
                UCF Data Analysis Program Configuration:

    Current Configuration:
    Readings/second  = 10   ------      New Trigger code (0-15): 0
    Run Duration     = 1.5 min --
    Input Channel    = CHO ------
    Maximum Range    = 10v ------
    Trigger Code     = 8    ------
    Relay 0 Start    = on   ------
    Relay 0 End      = off ------
    Relay 1 Start    = off ------
    Relay 1 End      = off ------
    I/O Address      = &H300 ----
    Interrupt Level  = 5    ------
    Graphics Adapter = EGAHIRES -
    Desktop Utility  = ----------
    c:\pctools\pcshell.exe
```

The interface system has been designed to get a trigger interrupt at IP3 alone (see Metrabyte manual for more information). The binary code for a single trigger from IP3 is 1000, which is 8 in decimal. Unless you use the board with another interface, this setting should be left at 8.

```
 IP3   IP2   IP1   IP0
  1     0     0     0    = 8
```

If you don't want a trigger (e.g. for a manual data acquisition run) set the trigger to 0.

Relay switches:

```
╔═══════════════════════════════════════════════════════════════╗
║            UCF Data Analysis Program Configuration:           ║
╠═══════════════════════════════════════════════════════════════╣
║  Current Configuration:                                       ║
║  Readings/second  = 10   ------                               ║
║  Run Duration     = 1.5 min --    ┌─────────────────────┐     ║
║  Input Channel    = CH0 ------    │  Relay 0 Start:     │     ║
║  Maximum Range    = 10v ------    │                     │     ║
║  Trigger Code     = 0    ------   │  F1:  ON            │     ║
║  Relay 0 Start    = on   ------   │  F2: OFF            │     ║
║  Relay 0 End      = off ------    │                     │     ║
║  Relay 1 Start    = off ------    │                     │     ║
║  Relay 1 End      = off ------    │                     │     ║
║  I/O Address      = &H300 ----    │                     │     ║
║  Interrupt Level  = 5    ------   │                     │     ║
║  Graphics Adapter = EGAHIRES -    │  DEFAULT = OFF      │     ║
║  Desktop Utility  = ----------    └─────────────────────┘     ║
║  c:\pctools\pcshell.exe                                       ║
╚═══════════════════════════════════════════════════════════════╝
```

Relay switches are not used for this interface, therefore all switches should be set to "off". These relays can be used to turn certain switches on the ultracentrifuge side on and off. They allow 2-directional communication between instruments. This interface is 1-directional only. If you have another instrument connected to the Metrabyte A to D card, you can use these switches to control the instrument, this program provides the software support for that, although it is not documented. If you have any special requirements, please contact:
Borries Demeler at P.O. Box 9252, Missoula, Montana 59807.

A-29

I/O address:

```
┌─────────────────────────────────────────────────────────────────┐
│░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░│
│░░░░░░░░░░░░░░░░░░░ UCF Data Analysis Program Configuration: ░░░░░░│
│░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░│
│                                                                   │
│  ┌──────────────────────────────────────────────────────────┐    │
│  │ Current Configuration:                                    │    │
│  │ Readings/second  = 10  ------                             │    │
│  │ Run Duration     = 1.5 min -- ┌──────────────────────┐    │    │
│  │ Input Channel    = CH0 ------ │  I/O Base Address:   │    │    │
│  │ Maximum Range    = 10v ------ ├──────────────────────┤    │    │
│  │ Trigger Code     = 0   ------ │ F1:  &H378 (LPT1)    │    │    │
│  │ Relay 0 Start    = off ------ │ F2:  &H2F0 (LPT2)    │    │    │
│  │ Relay 0 End      = off ------ │ F3:  &H3F8 (COM1)    │    │    │
│  │ Relay 1 Start    = off ------ │ F4:  &H2F8 (COM2)    │    │    │
│  │ Relay 1 End      = off ------ │ F5:  &H380 (SDLC)    │    │    │
│  │ I/O Address      = &H300 ---- │ F6:  &H200 (GAME)    │    │    │
│  │ Interrupt Level  = 5   ------ │ F7:  &H300 (CARD)    │    │    │
│  │ Graphics Adapter = EGAHIRES - │                      │    │    │
│  │ Desktop Utility  = --------- │ DEFAULT = &H300      │    │    │
│  │ c:\pctools\pcshell.exe        └──────────────────────┘    │    │
│  └──────────────────────────────────────────────────────────┘    │
└─────────────────────────────────────────────────────────────────┘
```

If you reconfigure your program by adding another device to your computer, you might have to change the interrupt level, the base address for your analog to digital board and the jumper switch settings on the board. Addressing memory locations in your computer is performed in hexadecimal format, since most PC's operate on a 16-bit basis. The address range for I/O operations is between &H200 - &H3FF (512 - 1023 decimal), with each device requiring a memory location, which is used as a start address for the memory operation. Following addresses are commonly used:

1) &H378 (decimal = 800) - LPT1 (Line printer 1)

2) &H2F0 (decimal = 752) - LPT2 (Line printer 2)

3) &H3F8 (decimal = 1016) - COM1 (RS 232 port 1)

4) &H2F8 (decimal = 760) - COM2 (RS 232 port 2)

5) &H380 (decimal = 896) - SDLC (SDLC communications)

6) &H200 (decimal = 512) - GAME (Game port)

7) &H300 (decimal = 768) - IBM prototype card

Other addresses are used for hard disks, memory expansions, binary communications, graphic monitors, floppy drives and others. For more information, refer to a computer users manual.

In order to determine, which address you want, find a device in the list which is not connected to your computer and use its address for the CROM-1 board. If you are not sure, number 7 (&H300) is usually a safe choice. You can then set the jumper switches on your CROM-1 board according to the address you choose. After you determine the base address you are going to use, view the switch setting info from the main menu.

Interrupt levels:

```
┌─────────────────────────────────────────────────────────────┐
│                                                             │
│            UCF Data Analysis Program Configuration:         │
│                                                             │
└─────────────────────────────────────────────────────────────┘
┌─────────────────────────────────────────────────────────────┐
│  Current Configuration:                                     │
│  Readings/second  = 10    ------                            │
│  Run Duration     = 1.5 min --    ┌──────────────────────┐  │
│  Input Channel    = CH0   ------  │   Interrupt Level:   │  │
│  Maximum Range    = 10v   ------  │                      │  │
│  Trigger Code     = 0     ------  │     F1: 2            │  │
│  Relay 0 Start    = off   ------  │     F2: 3            │  │
│  Relay 0 End      = off   ------  │     F3: 4            │  │
│  Relay 1 Start    = off   ------  │     F4: 5            │  │
│  Relay 1 End      = off   ------  │     F5: 6            │  │
│  I/O Address      = &H300 ----    │     F6: 7            │  │
│  Interrupt Level  = 5     ------  │                      │  │
│  Graphics Adapter = EGAHIRES -    │   DEFAULT = 5        │  │
│  Desktop Utility  = ----------    └──────────────────────┘  │
│  c:\pctools\pcshell.exe                                     │
└─────────────────────────────────────────────────────────────┘
```

I/O devices in a computer require interrupts, which allow the main processor to switch between devices. Your interrupt levels should be set not to conflict with other devices having the same interrupt level. Here are some common interrupt level settings:

1) interrupt 1 - used by system clock

2) interrupt 2 - reserved

3) interrupt 3 - used by COM2 RS 232 serial port if connected

4) interrupt 4 - used by COM1 RS 232 serial port if connected

5) interrupt 5 - used by hard disk or LPT2

6) interrupt 6 - used floppy disk drive adapter

7) interrupt 7 - used by LPT1 if connected

if you are not sure which interrupt level to use, 3 or 5 are usually a good choice.

Graphics adapter:

```
┌──────────────────────────────────────────────────────────────┐
│:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::│
│:::::::::::::::::::  UCF Data Analysis Program Configuration:  :::│
│:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::│
├──────────────────────────────────────────────────────────────┤
│                                                                │
│  Current Configuration:                                        │
│  Readings/second  = 10  ------                                 │
│  Run Duration     = 1.5 min --    ┌─────────────────────────┐  │
│  Input Channel    = CH0 ------    │  Graphics Adapter:      │  │
│  Maximum Range    = 10v ------    │                         │  │
│  Trigger Code     = 0   ------    │  F1: HERCULES           │  │
│  Relay 0 Start    = off ------    │  F2: IBM CGA            │  │
│  Relay 0 End      = off ------    │  F3: IBM EGA            │  │
│  Relay 1 Start    = off ------    │  F4: IBM EGA HIRES      │  │
│  Relay 1 End      = off ------    │  F5: IBM VGA            │  │
│  I/O Address      = &H300 ----    │  F6: IBM VGA MONO       │  │
│  Interrupt Level  = 5   ------    │  F7: IBM MONOCHROME     │  │
│  Graphics Adapter = EGAHIRES -    │                         │  │
│  Desktop Utility  = ----------    │  DEFAULT=VGA            │  │
│  c:\pctools\pcshell.exe           └─────────────────────────┘  │
│                                                                │
└──────────────────────────────────────────────────────────────┘
```

Choose the graphics adaptor and monitor you have connected to your computer. Do to the very high graphics requirements of the analysis programs, it is suggested to use a VGA-type monitor with this program.

SWITCH SETTINGS FOR THE CROM-1 BOARD:

If you reconfigure your program by adding another device to your computer, you might have to change the interrupt level, the base address for your analog to digital board and the jumper switch settings on the board. Addressing memory locations in your computer is performed in hexadecimal format, since most PC's operate on a 16-bit basis. The address range for I/O operations is between &H200 - &H3FF (512 - 1023 decimal), with each device requiring a memory location, which is used as a start address for the memory operation. Following addresses are commonly used:
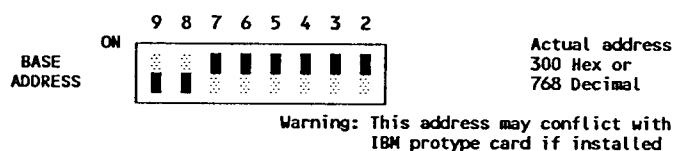
1) &H378 (decimal = 800) - LPT1 (Line printer 1)

2) &H2F0 (decimal = 752) - LPT2 (Line printer 2)

3) &H3F8 (decimal = 1016) - COM1 (RS 232 port 1)

4) &H2F8 (decimal = 760) - COM2 (RS 232 port 2)

5) &H380 (decimal = 896) - SDLC (SDLC communications)

6) &H200 (decimal = 512) - GAME (Game port)

7) &H300 (decimal = 768) - IBM prototype card

Other addresses are used for hard disks, memory expansions, binary communications, graphic monitors, floppy drives and others. For more information, refer to a computer users manual.

In order to determine, which address you want, find a device in the list which is not connected to your computer and use its address for the CROM-1 board. If you are not sure, number 7 (&H300) is usually a safe choice. You can then set the jumper switches on your CROM-1 board according to the address you choose:

Enter desired base address in decimal or &H--- form: &H300

Set switch sliders in positions shown below:-

```
                    9 8 7 6 5 4 3 2
               ON  ┌─────────────────┐      Actual address
   BASE            │ ░ ░ █ █ █ █ █ █ │      300 Hex or
   ADDRESS         │ █ █ ░ ░ ░ ░ ░ ░ │      768 Decimal
                   └─────────────────┘
                   Warning: This address may conflict with
                            IBM protype card if installed
```
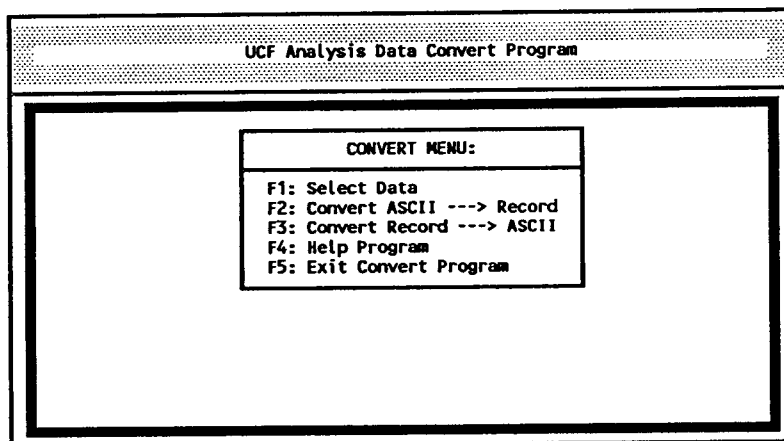
To determine the accurate switch setting, run the switch setting help program.

DESKTOP UTILITY:

If you have a desktop utility program, specify the drive, data path and full filename, including the extension, of your favorite desktop utility. This program is intended to help you organize the data on your disk. It should allow functions like editing, copying, deleting and other file management options. It is not supplied with this program, but it can be added with the configuration program utility, so chaining to the program is possible. To include this option, you should add the full path and filename (including extension) of the desktop utility you want to use, when prompted by the configuration program.

CONVERT PROGRAM:

The convert program allows you to convert datafiles from binary record format to ASCII format. ASCII formatted files can be imported into other programs, such as SIGMAPLOT, QUATTRO-PRO, LOTUS 1-2-3, SUPERCALC, etc. Initially, datafiles are saved as binary record files, which allows faster access to the disk and consumes less disk space than ASCII formatted files. All data I/O in the ULTRASCAN PROGRAM is performed in this binary record format.

```
                    UCF Analysis Data Convert Program


             ┌─────────────────────────────────────┐
             │            CONVERT MENU:             │
             ├─────────────────────────────────────┤
             │  F1: Select Data                     │
             │  F2: Convert ASCII ---> Record       │
             │  F3: Convert Record ---> ASCII       │
             │  F4: Help Program                    │
             │  F5: Exit Convert Program            │
             └─────────────────────────────────────┘
```

To use the convert program, run "SELECT DATA" and select the run identification and the directory and datapath, where the datafiles are located. Select the desired function and then select the cells and scans from the run you want to convert. The program will then perform the conversion and tell you about the status of the conversion. Remember, that you have to reconvert ASCII formatted files into record files, before you can use them in the ULTRASCAN PROGRAM.

ROTOR INFORMATION PROGRAM:

With this program all the information for rotor calibration, RTIC values and temperature conversion can be handled.

```
 ....................................................................
 :::::::::::::::::::  Ultracentrifuge Rotor Information Program  ::::::::::
 ....................................................................

         ┌──────────────────────────────┐
         │          ROTOR MENU:          │
         ├──────────────────────────────┤
         │ F1: Add new Rotor to List     │
         │ F2: Calibrate new Rotor       │
         │ F3: View Calibration Data     │
         │ F4: Delete Rotor from List    │
         │ F5: Calculate RTIC            │
         │ F6: Calculate Temperature     │
         │ F7: Help Program              │
         │ F8: Exit Rotor Program        │
         └──────────────────────────────┘
```

This program will update the rotor list and calibration table for each rotor. To enter new rotor information, first enter the name for the new rotor, using "Add new Rotor to List". Simply enter the rotor name, i.e.: AN-F 322. Then you can enter the calibration data by selecting "Calibrate new Rotor":

Calibration Program for Ultracentrifuge Rotors:

```
┌──────────────────────┐
│    Select Rotor:     │
├──────────────────────┤
│ F1: AN-F   655       │
│ F2: AN-F   398       │
│ F3: AN-J   188       │
│ F4: AN-D  3376       │
│ F5: AN-K   367-SP    │
└──────────────────────┘
```

You will be prompted for the rotor you want to calibrate (you can select an existing rotor for re-calibration as well) and then for temperature and RTIC values.

Calibration Program for Ultracentrifuge Rotors: Rotor = AN-F 655

Please enter the calibration data (enter 'q' if you are done):

Temperature for 1. data point: 23.4
RTIC value for 1. data point: 5402

When finished, enter "Q" to quit and the data will be fitted to a straight line by a least squares fitting routine. You will see the fitted data displayed as an equation of the format:

Temperature = a * RTIC + Zero-temperature

A-36

Any rotor may be deleted using "Delete Rotor from List". Using any of the calculation functions, you can enter a temperature to find out the corresponding RTIC value or vice-versa. Select "View Calibration Data" to look at currently defined rotor calibrations and fitted equations.

```
This is the Rotor Information for all Rotors:

Rotor 1: AN-F  655:    Temperature = -.020789 * RTIC Value + 131.79

Rotor 2: AN-F  398:    Temperature = -.02055 * RTIC Value + 132.881

Rotor 3: AN-J  188:    Temperature = -.020754 * RTIC Value + 137.647

Rotor 4: AN-D 3376:    Temperature = -.02081 * RTIC Value + 135.888

Rotor 5: AN-K 367-SP:   Temperature = -.02138 * RTIC Value + 135.516

Hit any key to continue...
```

Entering inner reference measurements:

For each rotor the program assumes that a different reference cell exists. Although differences may be small, it is possible to enter precise rotor measurements in this program. The radial distance from the center of the rotor axis to the center of the reference cell according to Beckman is 65 mm for all rotors, therefore it is assumed to be correct. However, the measurements for the reference cells can be modified if the reference cells are measured in a machine shop or a with a micro-comparator. This data is entered in the rotor program (see: UTILITY PROGRAM) and used for the calculation of the cell length correspondent (see: Edit Cell Length Correspondent).

EDIT PROGRAM:

The first menu in the edit program lets you choose between various edit functions.

```
UCF Analysis Data Edit Program

                    EDIT MENU:

            F1: Select Data
            F2: View Data
            F3: Edit Sedimentation Velocity
            F4: Edit Sedimentation Equilibrium
            F5: Utilities
            F6: Help
            F7: Exit Edit Program
```

First, you select the data you want to edit with the "Select Data" function. By default, this is set to the last run you made with the data acquisition program. If different, simply enter the name of your run identification and the directory name and path where the data is stored. The first thing to do is to view a plot of the raw data, to get an impression of the data accumulated, so select "View Data" and then from the next menu, select "View raw Data":

```
Please select which data you want to view:

                    VIEW MENU:

            F1: View raw Data
            F2: View edited Data
            F3: Help
            F4: Exit View Program
```

If the data has not been edited, only the raw data can be viewed. You can select to view only a single scan or a subset of scans. For example, you might choose to view only the scans from the second of three cells. This data contains the outer and inner reference and data surrounding the reference holes:

Cell Nr. 2, Run Nr. 7   Wait, loading data...
Datafile: D:\DATA\50497.RC2
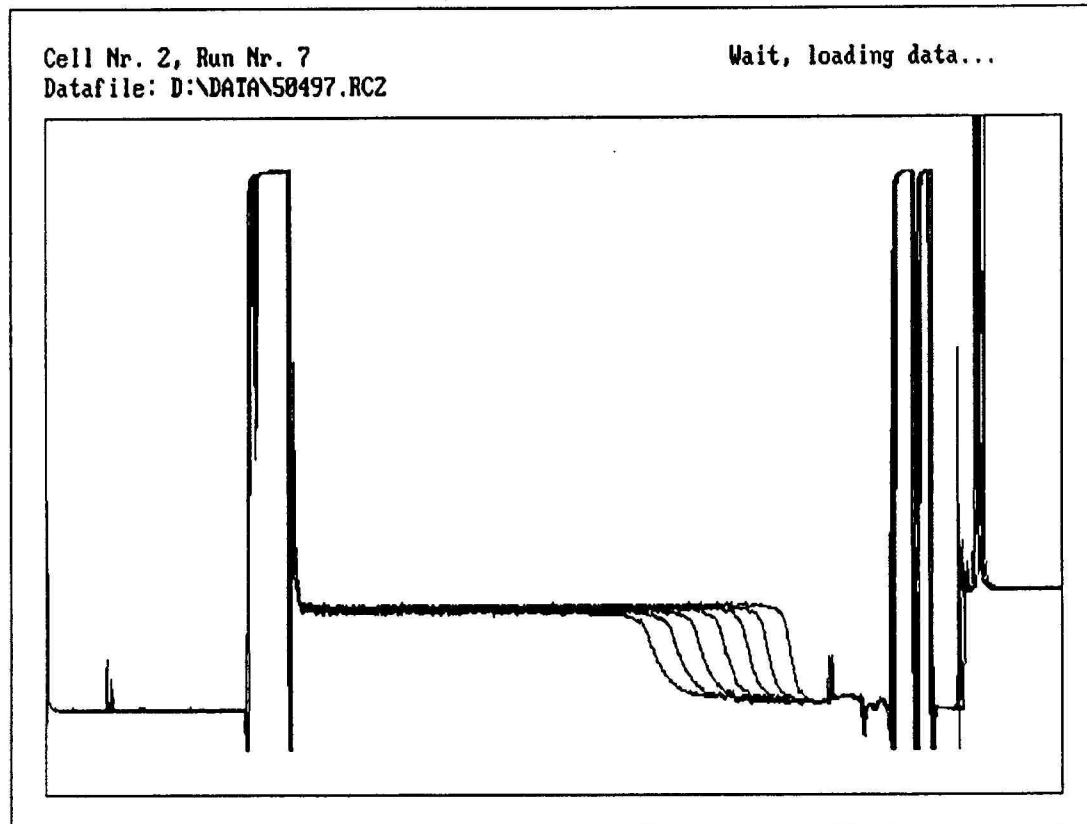
After displaying the data on the screen, you have the option to send a graphics screen dump to the lineprinter, provided the screen driver (usually GRAPHICS.COM from DOS) is loaded into memory, and the printer is connected. Simply answer "y" if you want to print out the data.

EDITING SEDIMENTATION VELOCITY DATA

To create edited data, select the appropriate edit function from the main menu of the edit program. You will be prompted for various actions with the mouse. The program guides you through the editing steps required. It will automatically present all scans specified from one experiment in sequential order. To give precise information, you can zoom the plot by defining an upper and lower limit to be viewed on the entire screen. First, you have to decide if any scans should be excluded from the analysis. Use your judgment and discard scans which are obviously containing noisy or otherwise useless data:


```
Cell Nr. 1, Run Nr. 7
Datafile: D:\DATA\50497.RC1        Please enter scans to be excluded: 1. 5
```
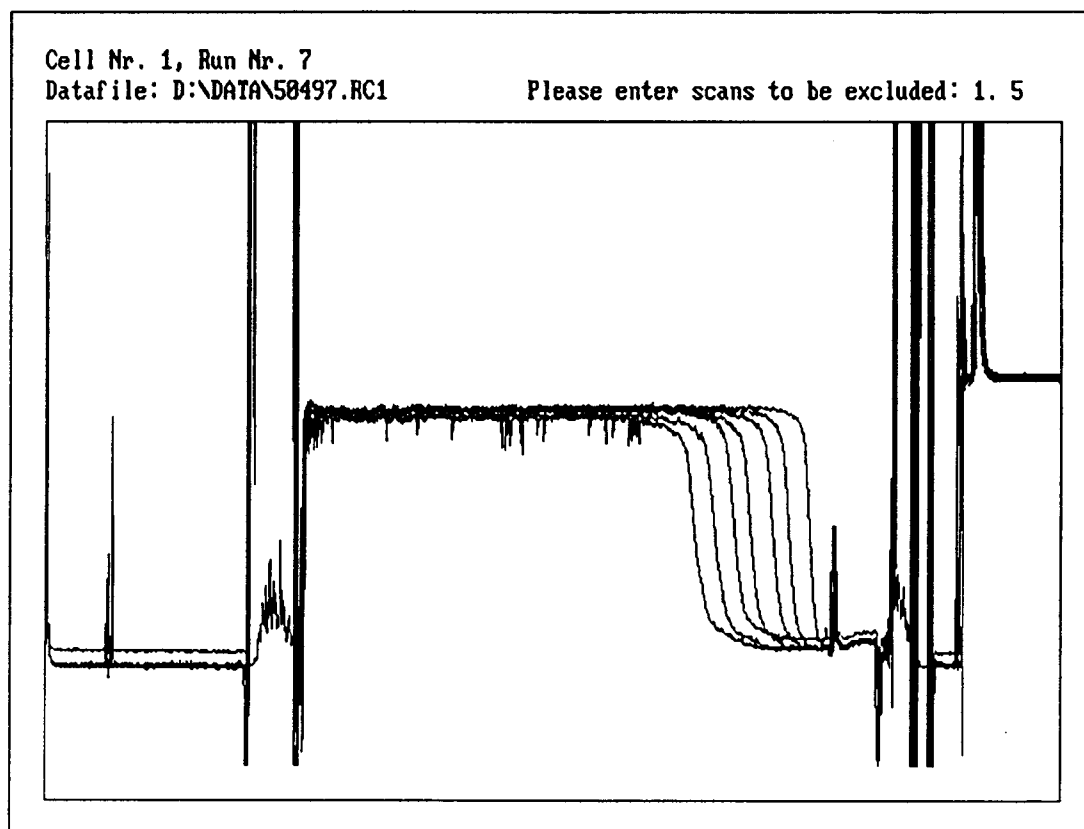
In this example, scan #5 should be excluded, since its baseline is different from the base lines of the other scans.

Then you select a scan for the midpoint definition. this is only done for a single scan. Therefore, select a scan where a midpoint can be easily defined. If more than one midpoint exists (because of multiple components) select the scan which contains the best information for the component of highest concentration. The corresponding midpoints of the other scans will be defined automatically.

```
Cell Nr. 1, Run Nr. 7
Datafile: D:\DATA\50497.RC1        Select Run Nr. for midpoint selection: 7
```
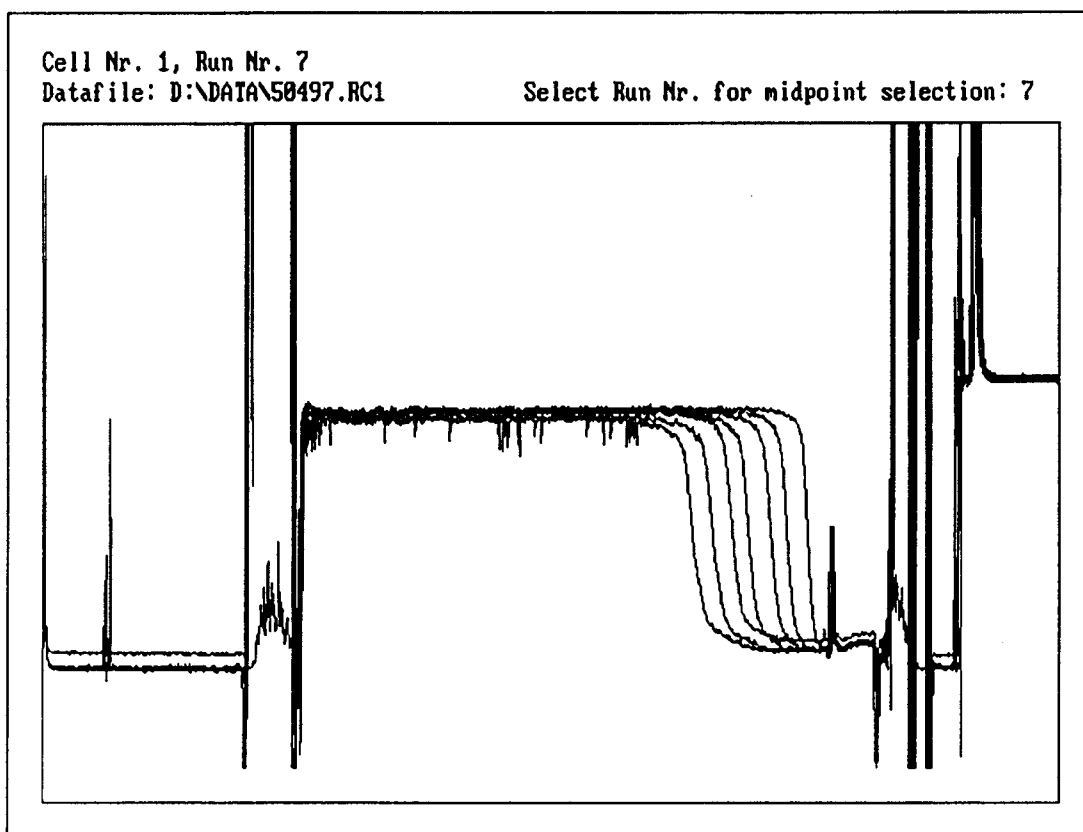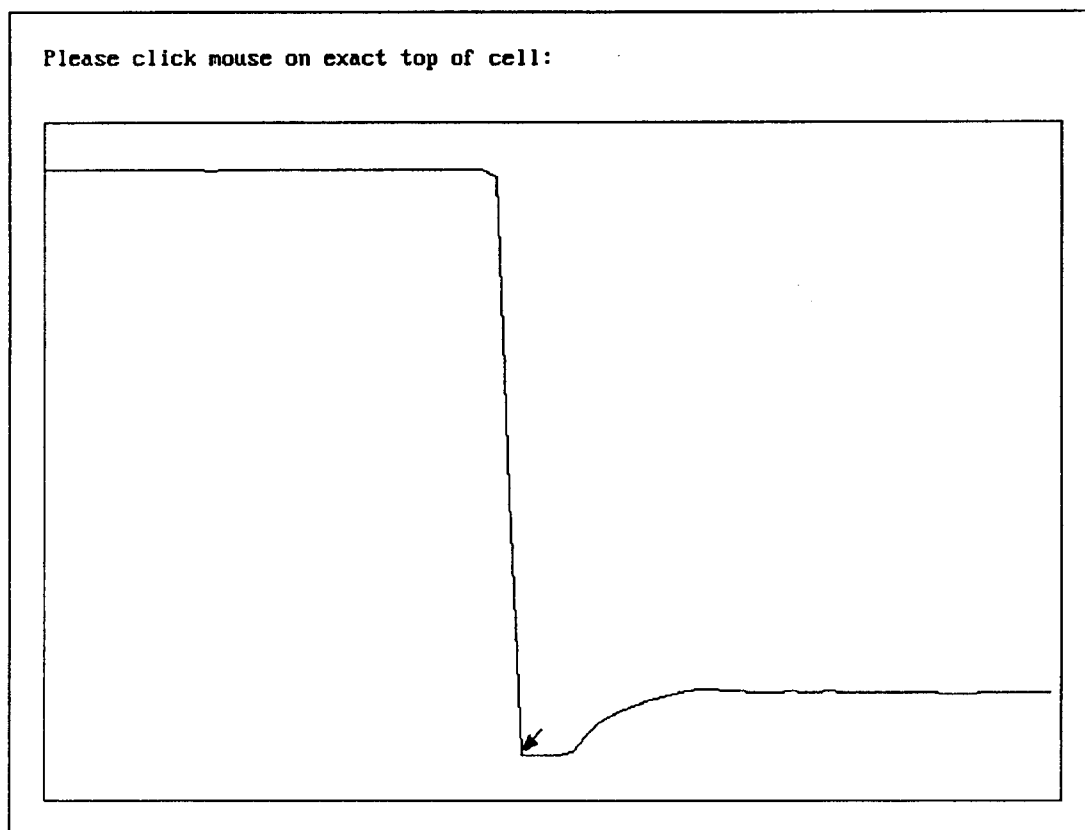
Next, you will define an average for the upper boundary and the baseline, then zoom around the midpoint and define the exact midpoint. You can also look at the derivative, if the data is not too noisy or the resolution is not too high (> 10 data points/second) to help you to define the midpoint. The largest downward peak in the derivative should point to the midpoint. However, if for some reason your data is noisy or the data is too closely spaced, the numerically determined derivative will not be of much use. Also, late scans have a more spread out boundary due to diffusion, which makes the derivative less discerning.


Please click mouse on exact midpoint of scan:
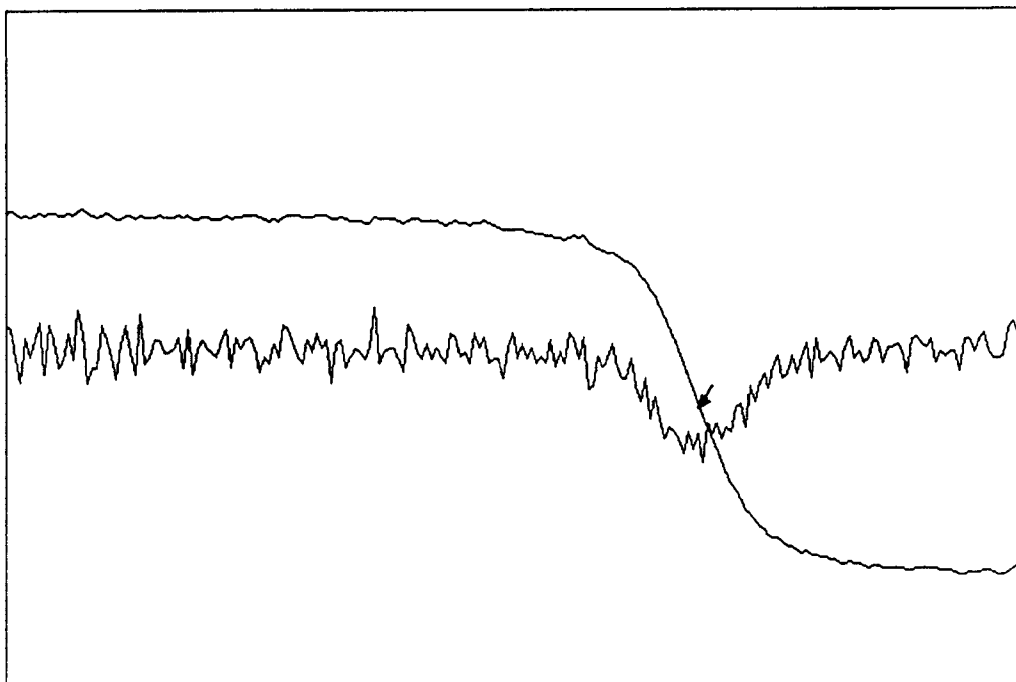
In the first scan to be edited from each cell, you have to define the top of the cell by zooming and clicking on the upper reference. The top of the cell will always be on the right side of the screen:

Please click mouse on exact top of cell:

Also, you will have to define the upper average (for the radial dilution effect) and the meniscus. The upper average is determined by clicking with the mouse around a linear region in the plateau without much noise. The data points enclosed in this region will be summed and averaged. For all subsequent scans only the upper average and the meniscus need to be defined. Again, zooming allows you to precisely determine the meniscus:
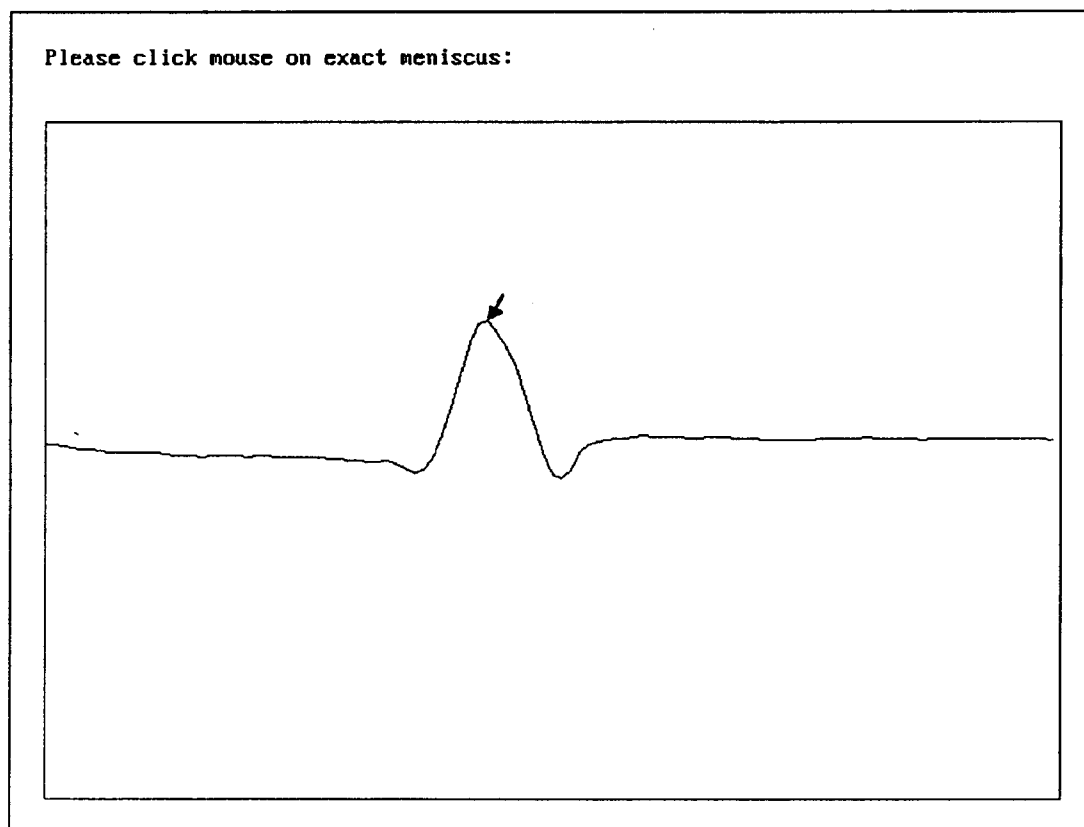

Make sure you use the meniscus of the sample and not the meniscus of the buffer. If you loaded the cells properly, the left meniscus, pointing upward, is the meniscus belonging to the sample.

After editing each cell you have the option to back up in case of a mistake:

```
Continue with next Scan?    YES      NO
```

When you are finished with the editing, all scans from a cell are plotted in an overlay fashion, so you can determine if the scans are well aligned and didn't leak during the run:

```
Cell Nr. 1
This is the edited data, press any key to continue...
```
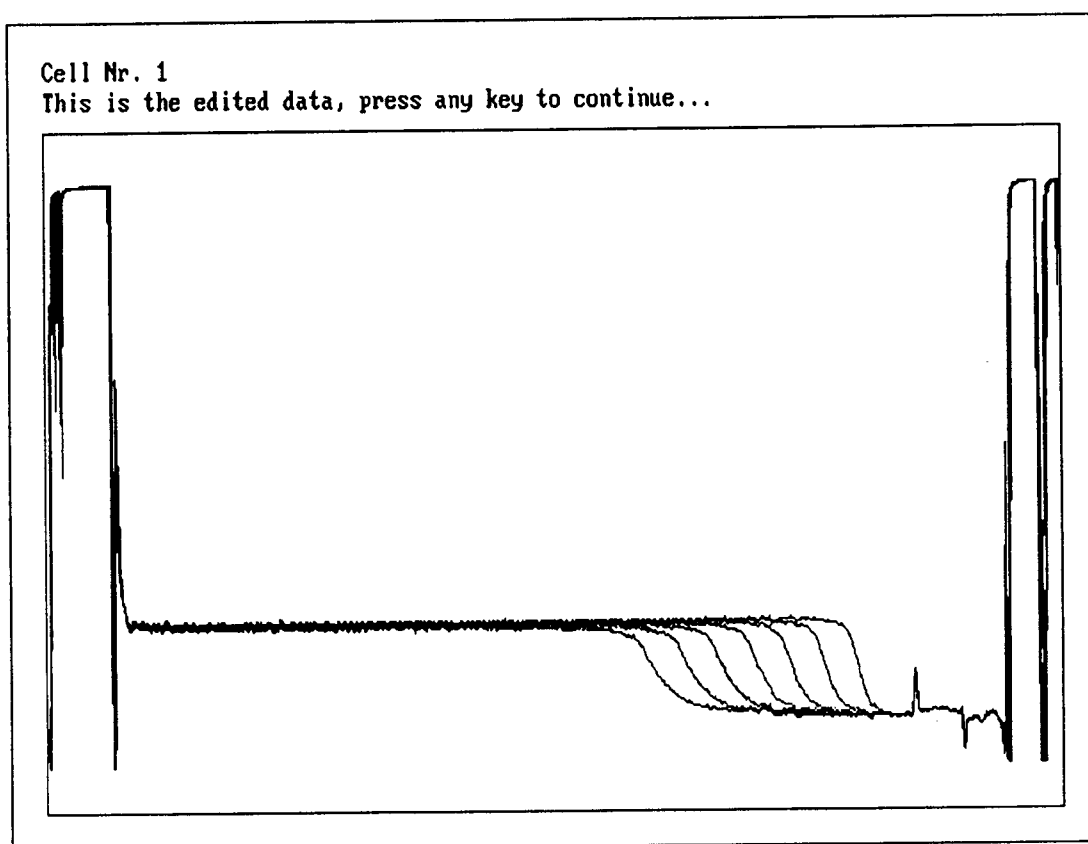

If they are not well aligned, you have to start over again. Also, if your cells leaked, the bottom reference marks will not be aligned, but shifted against each other (then your data is worthless and the experiment needs to be repeated). Remember, proper alignment is crucial for exact data analysis. All edited data are saved in a separate file, with extension ".E" + cell number. To inform you of the scan number you are working on, the name of the data file is printed at the top of the screen.

EDITING SEDIMENTATION EQUILIBRIUM DATA:

After selecting "Edit Sedimentation Equilibrium" from the EDIT MENU, all selected scans will be displayed on the screen. At this point you can exclude noisy or otherwise useless scans from the analysis. First, enter how many scans are to be excluded, and then the scan number of the excluded scan:

```
Cell Nr. 1, Run Nr. 4
Datafile: D:\DATA\50854.RC1        Do you want to exclude any scans? (y/n) _
```
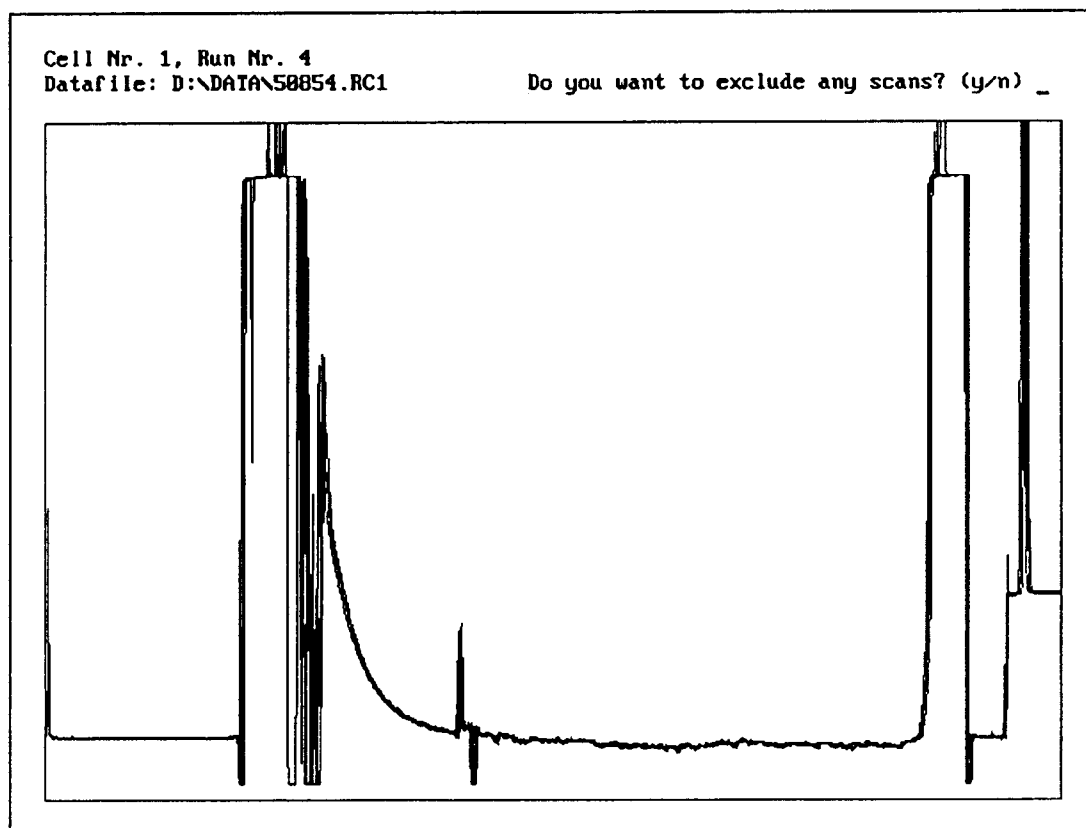
If no scans are to be excluded, enter "n" and continue by selecting a scan for the estimation of the baseline, usually the first scan:


```
Cell Nr. 1, Run Nr. 1
Datafile: D:\DATA\50851.RC1
Please use mouse to define left limit to average baseline absorbance:
```
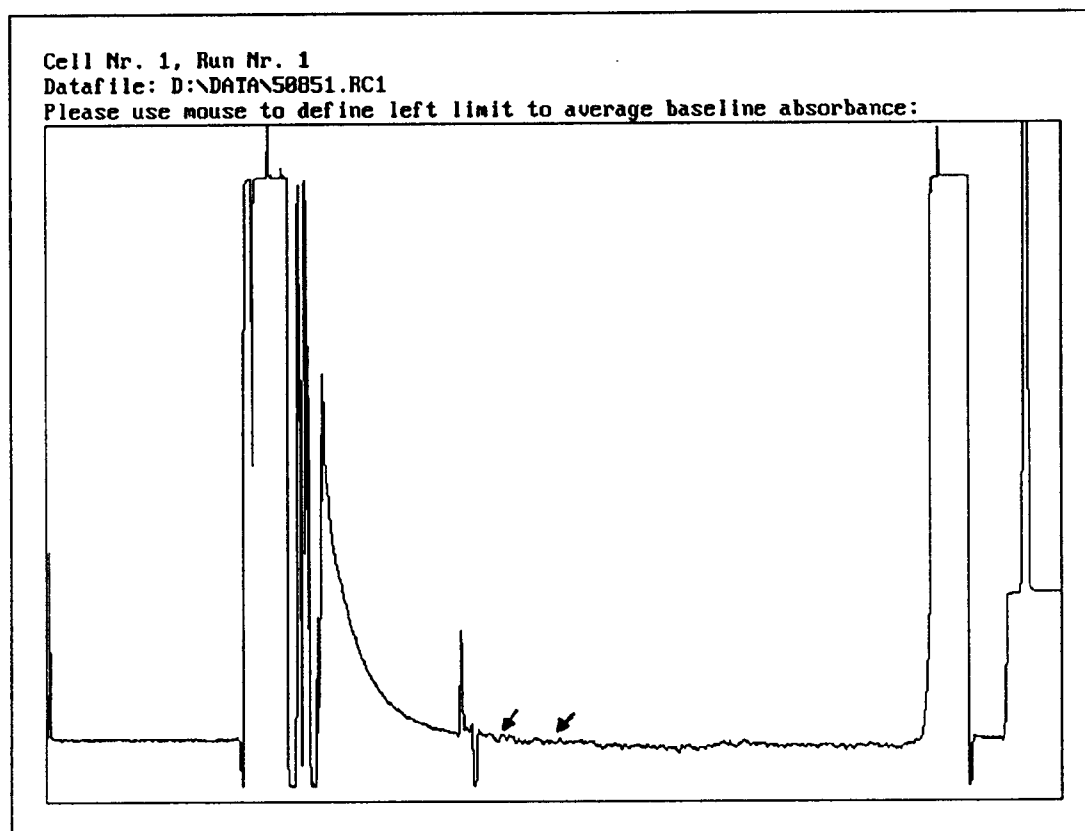
At this point you can set a left and right limit in the baseline among which the data will be summed and averaged. This average will be used as a starting estimate for the true baseline, which will be calculated by a chi-square minimization.

A-47

The next step is to define an upper limit for the largest absorbance data still to be included in the analysis:

```
Cell Nr. 1, Run Nr. 1                          Wait, loading data...
Datafile: D:\DATA\50851.RC1
Please use mouse to define left border of X-zoom around upper limit:
```
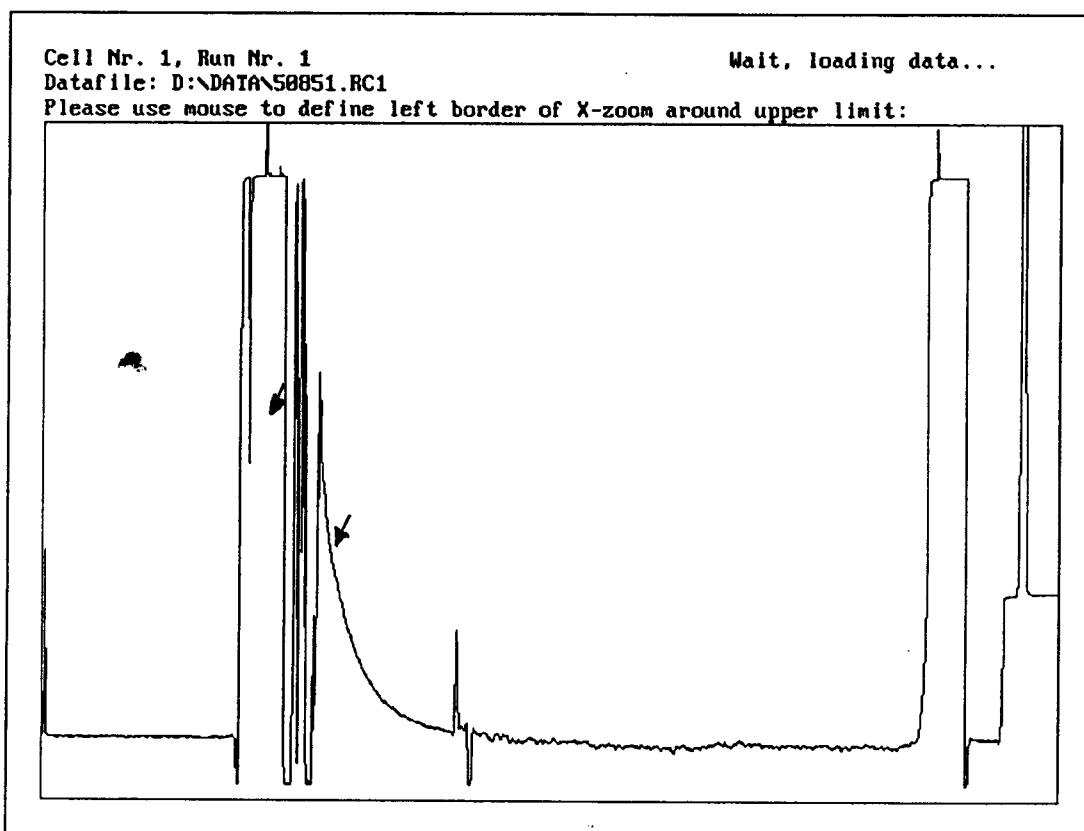
By providing a left and right limit, you can zoom this point and click with the mouse on the exact point to be used as a cutoff:

```
Please click mouse on exact upper limit:          Wait, loading data...
```
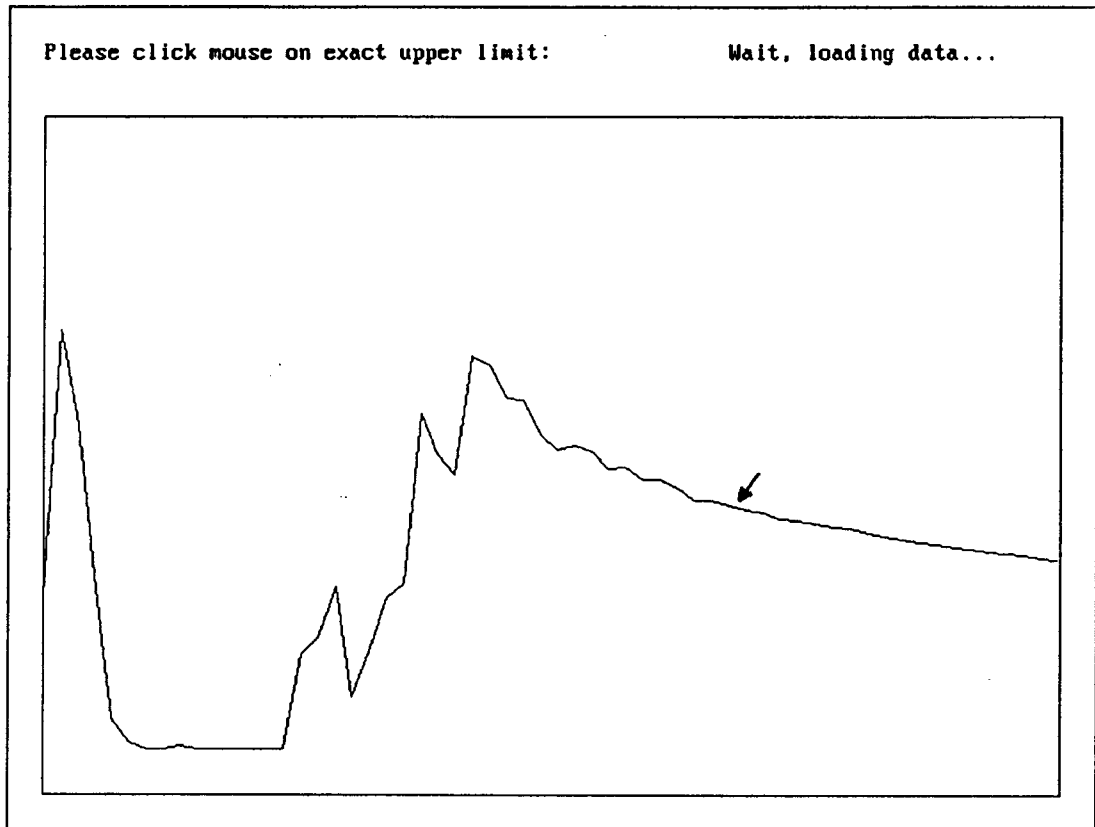
The next step is to define the top of the cell, the inner reference. This point can again be zoomed by providing a left and right limit by clicking with the mouse:

Cell Nr. 1, Run Nr. 1
Datafile: D:\DATA\50851.RC1
Please use mouse to define left border of X-zoom around top of cell:
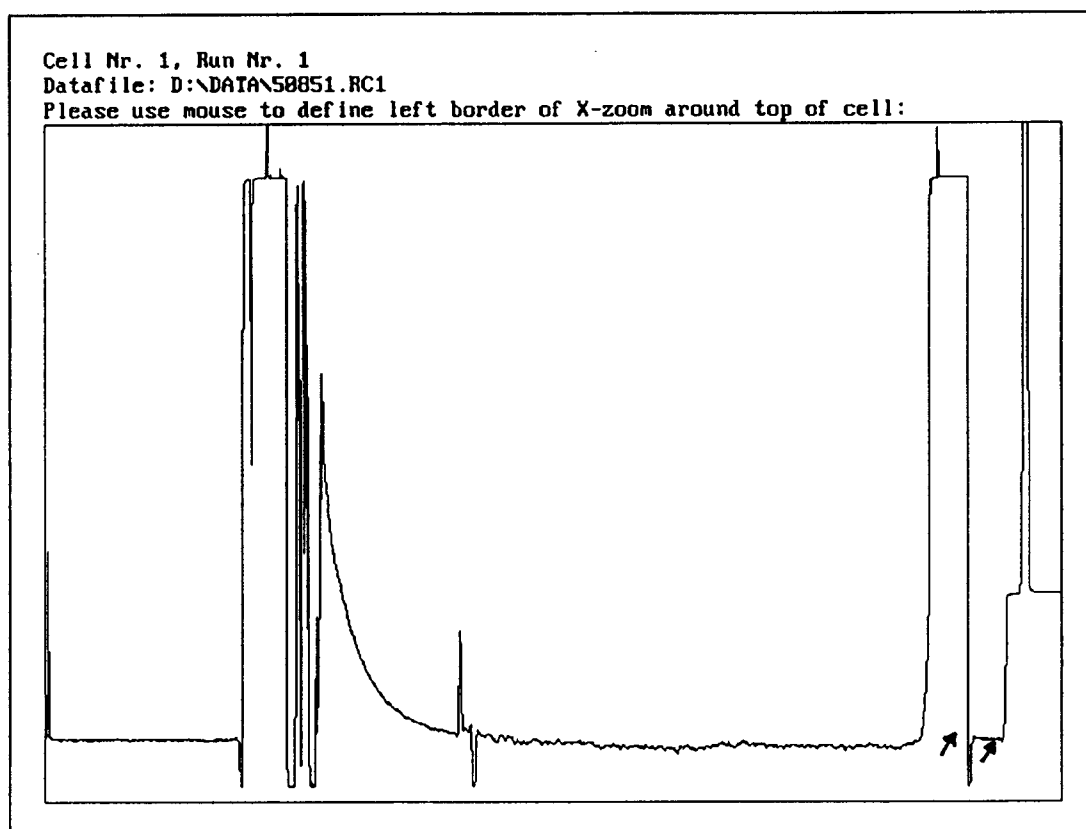
Please click mouse on exact top of cell:
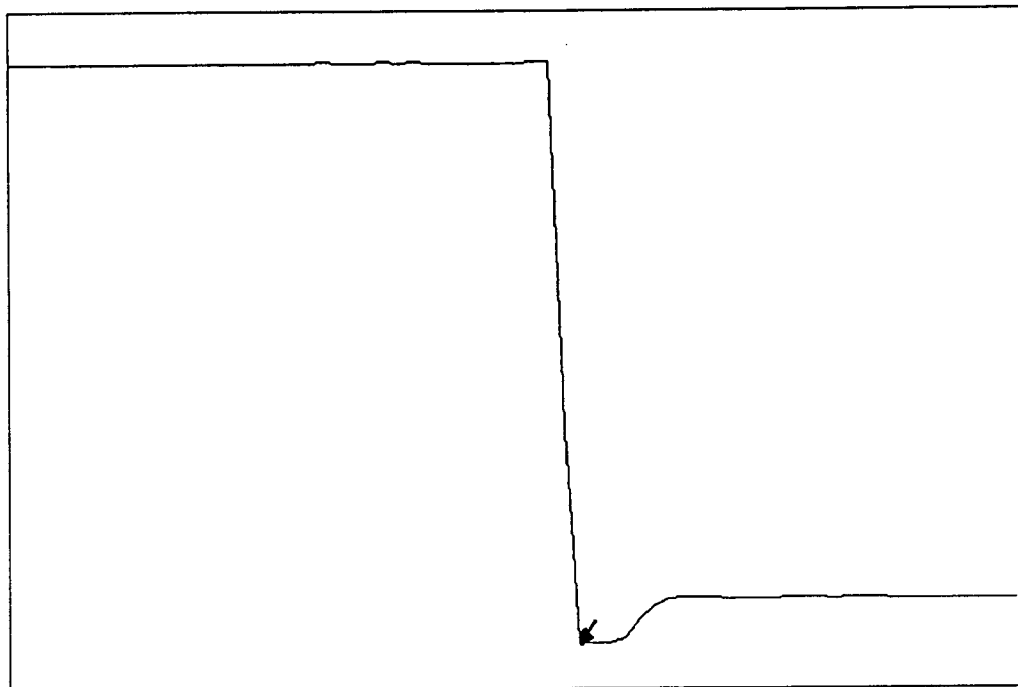
Click with the mouse on the point indicated.

To define the meniscus, zooming around the meniscus allows exact determination of the true meniscus point:

```
Cell Nr. 1, Run Nr. 1
Datafile: D:\DATA\50051.RC1
Please use mouse to define left border of X-zoom around meniscus:
```
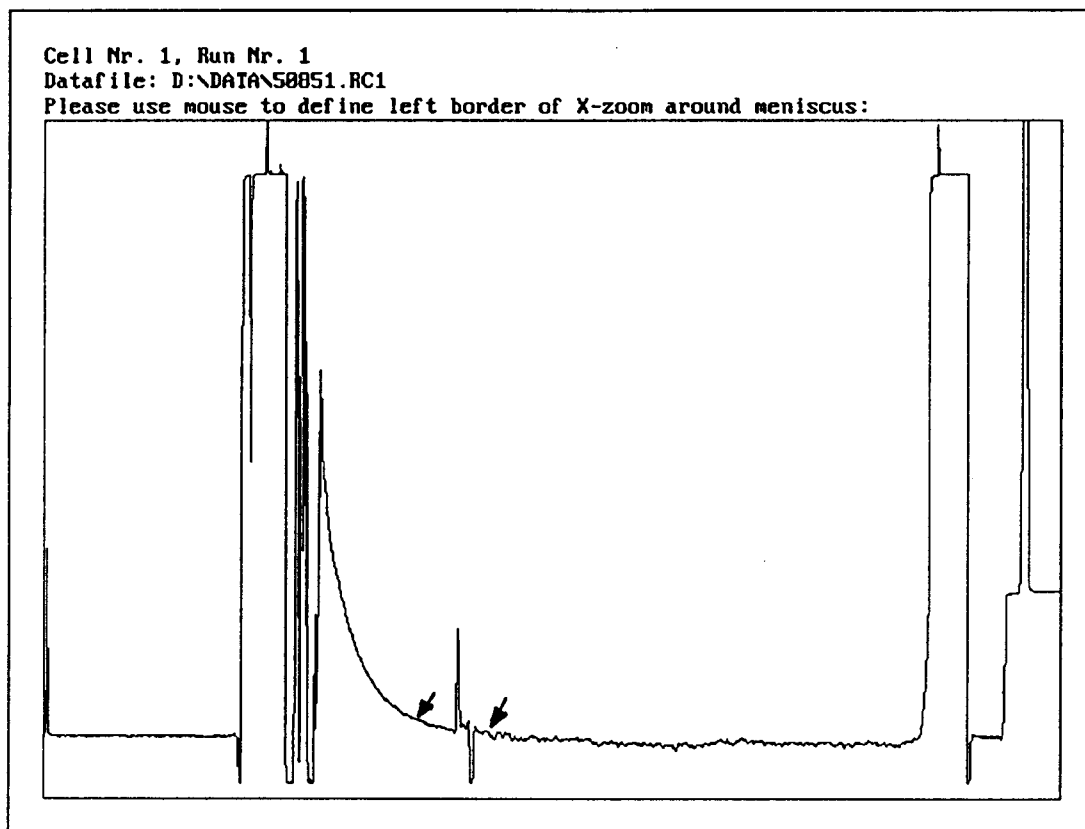
```
Please click mouse on exact meniscus:
```
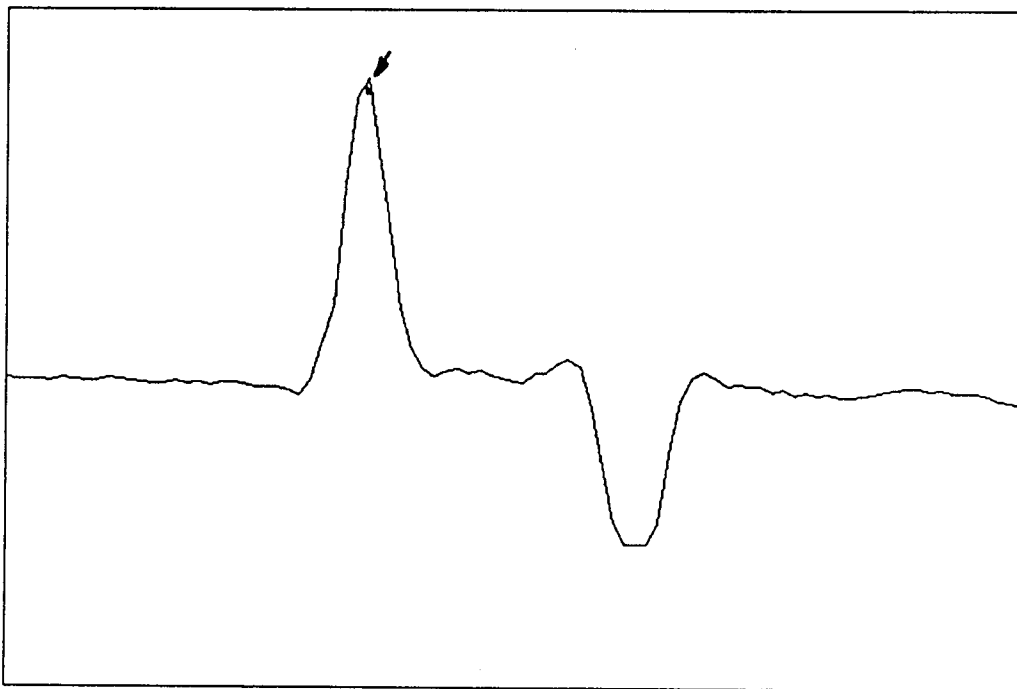

Click on the meniscus produced by the interface between buffer and sample, not on the meniscus resulting from the interface between buffer and air. The correct meniscus will be pointing upwards.

Continue editing each scan by defining the meniscus of each equilibrium scan included in the analysis. If you made a mistake, you have the option to back up and re-edit a scan:

```
Continue with next Scan?    YES    NO
```

When you are finished with the editing, all scans from a cell are plotted in an overlay fashion, so you can determine if the cells leaked:

```
This is the edited data of Cell #1, press any key to continue...
```
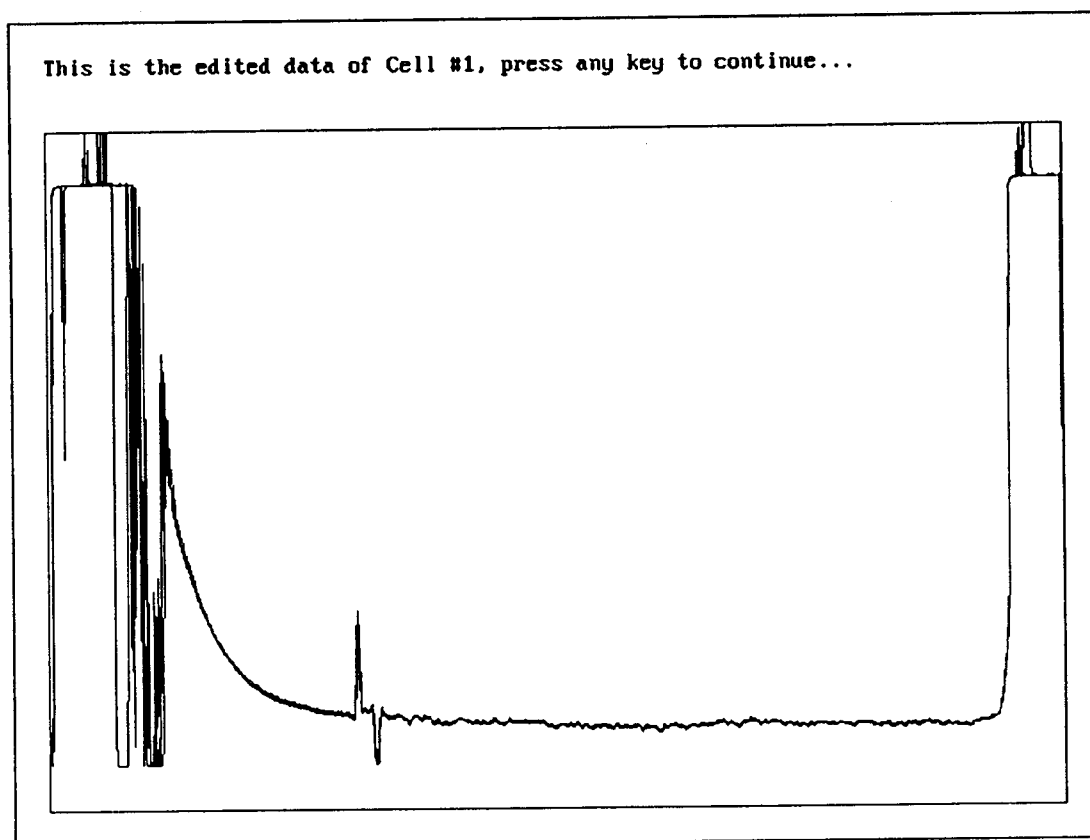

If the run has reached equilibrium, the traces of each scan should overlay exactly. If the trace is thicker at either end of the scan, equilibrium has not been reached, at least in the earlier scans. This provides a test for the status of the run.

EDITING THE CELL LENGTH CORRESPONDENT:

For each rotor the program assumes that a different reference cell exists. Although differences may be small, it is possible to enter precise rotor measurements in this program. The radial distance from the center of the rotor axis to the center of the reference cell according to Beckman specifications is 65 mm for all rotors, therefore it is assumed to be correct. However, the measurements for the reference cells can be modified if the reference cells are measured in a machine shop or with a micro-comparator. This data is entered in the ROTOR PROGRAM (see UTILITY PROGRAMS) and is used for the calculation of the of length correspondent. This is a measure for the distance between the inner and outer reference marks of each cell, normalized with respect to the number of data points taken between the two references.

In order to gain high accuracy with this measurement, it is recommended to use 10 individual scans acquired at the same resolution with damping/noise suppression on the interface and the stripchart recorder turned off. When you edit the cell length correspondent, the specified scans will be used for an average measurement of the number of data points between the inner and outer references. To the resolution provided by the data point density, the program will calculate a cell length correspondent based on the physical cell length (from the ROTOR PROGRAM) and the average number of data points.

It is recommended to run the cell length correspondent program each time anew when the scanner speed is modified or the number of data points/second setting is changed in PROGRAM CONFIGURATION. Also, if different reference cells are used, the program should be executed before the first analysis. Any acquired data can serve for the cell length correspondent determination.

To start the cell length correspondent program, select the run, number of cells and scans to be included and zoom around the top of the cell to specify the exact reference marks:

Cell Nr. 1, Run Nr. 1
Please use mouse to define left border of X-zoom around top of cell:

Click with the mouse on the indicated point.

Please click mouse on exact top of cell:
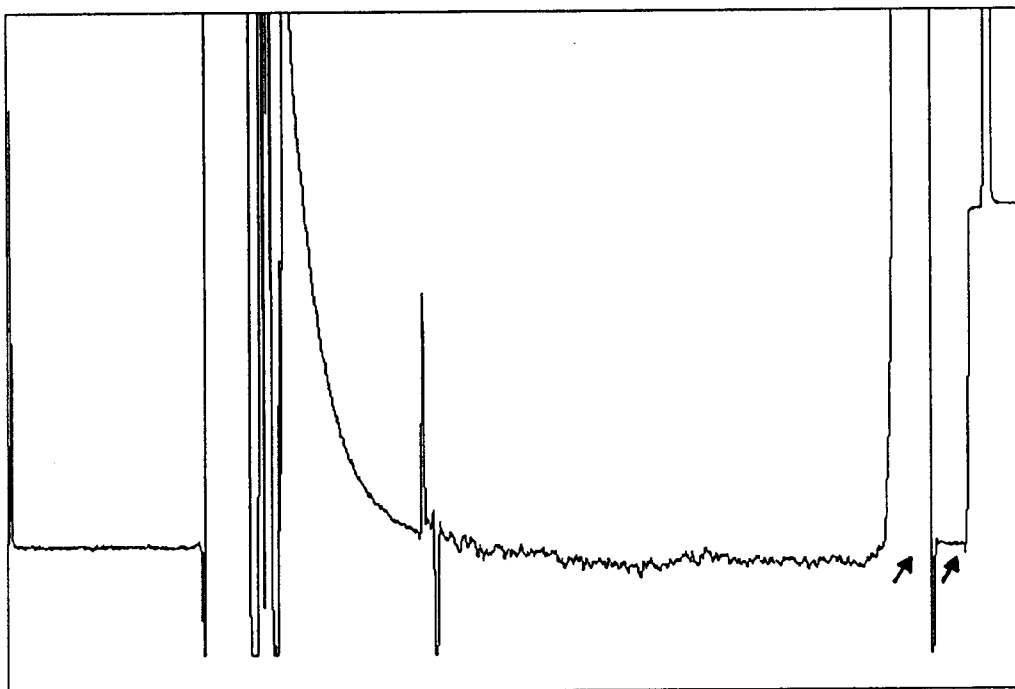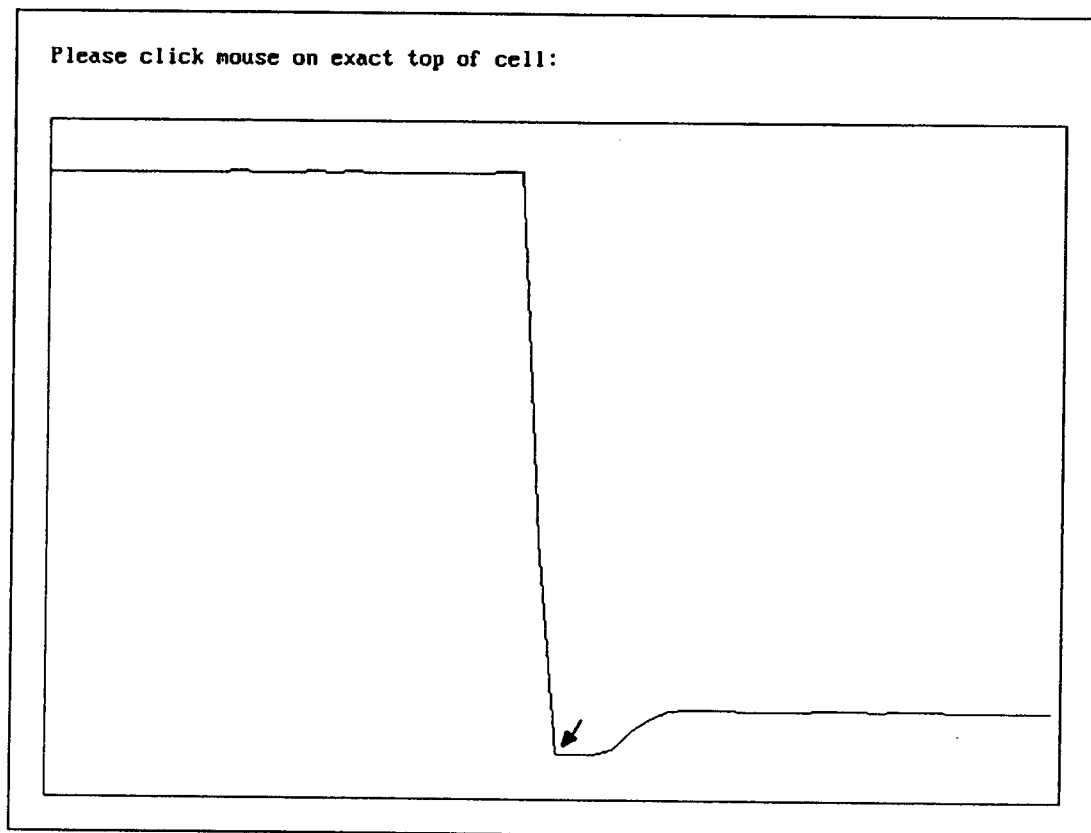
Now zoom around the bottom of the cell (outer reference mark):

Cell Nr. 1, Run Nr. 1
Please use mouse to define left border of X-zoom around bottom of cell:
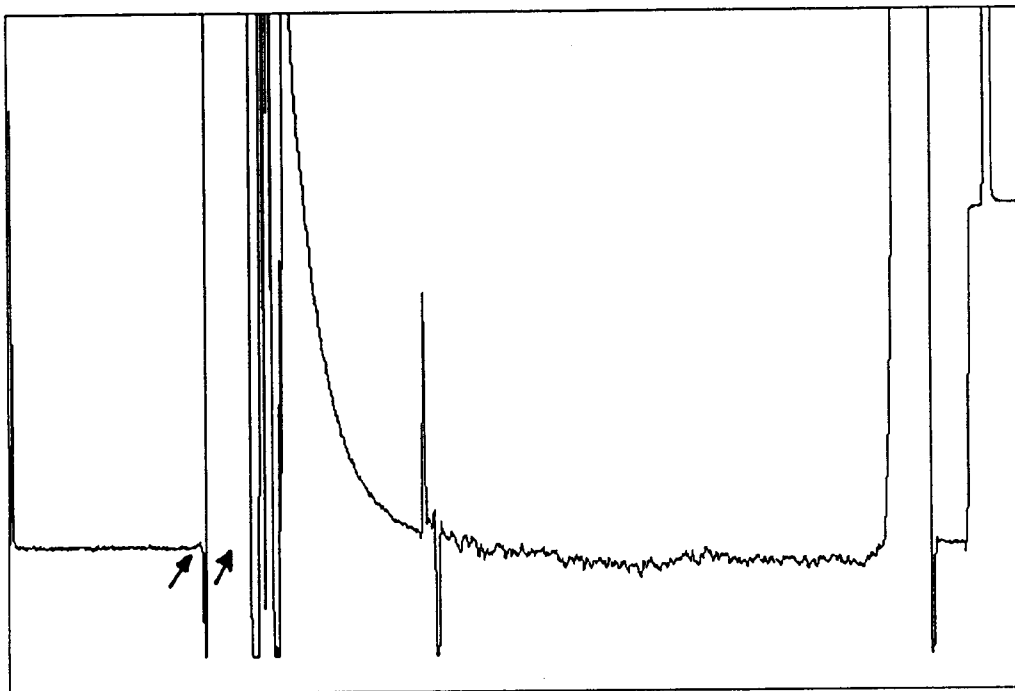
Click with the mouse on the indicated point. If you made a mistake, you can back up and re-edit the scan by clicking on the "NO" box:

Continue with next Scan?   YES   NO

Please click mouse on exact bottom of cell:
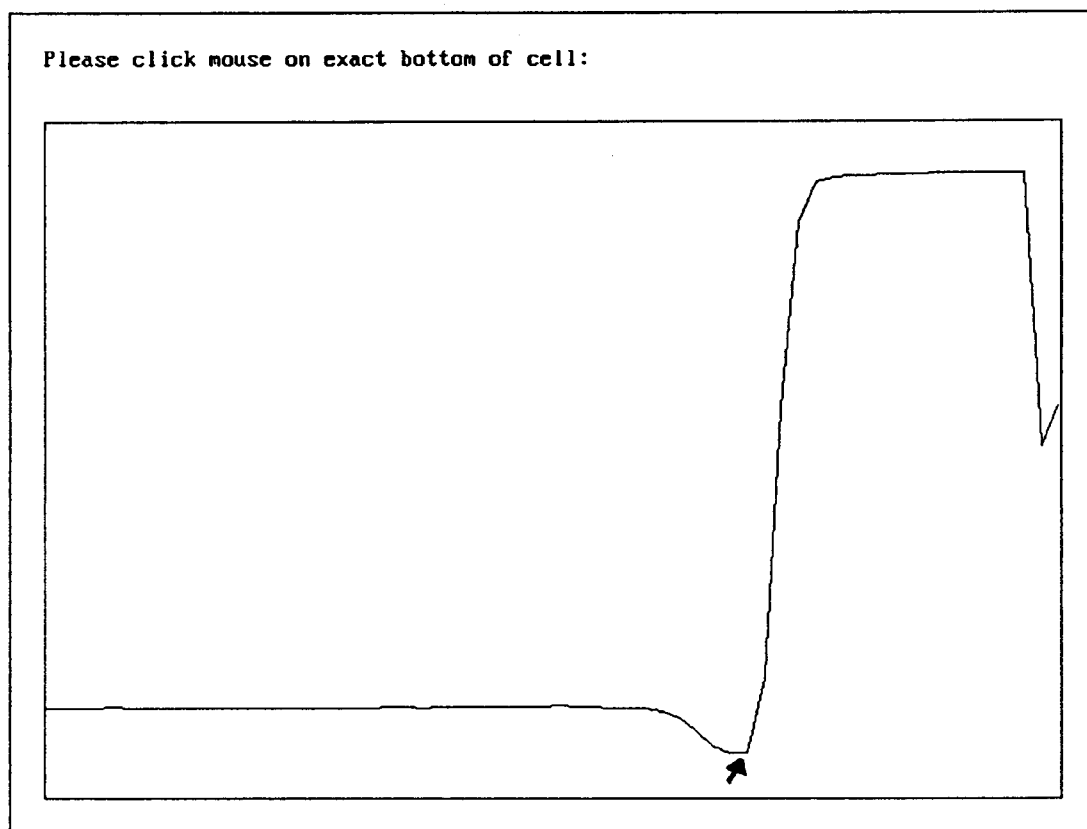
The last screen will show you the results and the new cell length correspondent:

There were 20 data points/second accumulated in Run TEST01.

```
Number of Data points for Cell # 1, Scan # 1 1291
Number of Data points for Cell # 1, Scan # 2 1289
Number of Data points for Cell # 1, Scan # 3 1290
Number of Data points for Cell # 1, Scan # 4 1290
Number of Data points for Cell # 2, Scan # 1 1291
Number of Data points for Cell # 2, Scan # 2 1290
Number of Data points for Cell # 2, Scan # 3 1289
Number of Data points for Cell # 2, Scan # 4 1290
```

This is the calculated average for all cells and scans: 1290

This is the new Cell Length Correspondent: 64.5

Hit any key to continue...

The new information will be written to a file called "ROTOR.DAT" where it will be used in the editing and analysis programs.

ANALYSIS PROGRAMS:

With the analysis programs edited data can be analyzed. The analysis program menu has several functions:

```
┌─────────────────────────────────────────────────────────────┐
│ ▓▓▓▓▓▓▓▓▓▓▓▓▓ Ultracentrifuge Data Analysis Program: ▓▓▓▓▓▓▓ │
│ ▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓ │
│ ┌─────────────────────────────────────────────────────────┐ │
│ │        ┌──────────────────────────────────────┐         │ │
│ │        │            ANALYSIS MENU:             │         │ │
│ │        ├──────────────────────────────────────┤         │ │
│ │        │ F1: van Holde - Weischet Analysis     │         │ │
│ │        │ F2: Second Moment Analysis            │         │ │
│ │        │ F3: Equilibrium Analysis              │         │ │
│ │        │ F4: Set S(W,20) Corrections           │         │ │
│ │        │ F5: Help                              │         │ │
│ │        │ F6: Exit Program                      │         │ │
│ │        └──────────────────────────────────────┘         │ │
│ │                                                         │ │
│ └─────────────────────────────────────────────────────────┘ │
└─────────────────────────────────────────────────────────────┘
```

VAN HOLDE - WEISCHET ANALYSIS:

Selecting the van Holde - Weischet analysis allows to do multi-component boundary analysis and sedimentation coefficient determination. After selecting the data, you have to select the hydrodynamic parameters to make accurate temperature and buffer corrections:

```
╔════════════════════════════════════════════════════════════╗
║            Selection of Hydrodynamic Parameters:           ║
╠════════════════════════════════════════════════════════════╣
║  ┌──────────────────────────────────────────────────────┐  ║
║  │                                                      │  ║
║  │  Current Selection:                                  │  ║
║  │                                                      │  ║
║  │  This is for Run ID 5083 at 19.12 °C:                │  ║
║  │                                                      │  ║
║  │  1.  V-BAR for your Sample:              .730        │  ║
║  │  2.  Density of your Buffer at 19.12 °C: 1.040       │  ║
║  │  3.  Viscosity of your Buffer at 19.12 °C: 1.112     │  ║
║  │                                                      │  ║
║  │  Do you want to change these entries (y/n)?          │  ║
║  │                                                      │  ║
║  │                                                      │  ║
║  │                                                      │  ║
║  │                                                      │  ║
║  └──────────────────────────────────────────────────────┘  ║
╚════════════════════════════════════════════════════════════╝
```

Just enter 'y' if you want to change an entry and select the number of the item to be changed. You will be prompted for the new value. After that, select the range of cells and scans to be included in the analysis and the number of integral divisions for the Weischet analysis. The maximum number of integral divisions is 50. However, if the data resolution is too low, not all 50 divisions can be made. An error will occur and require less divisions to be used. Following window will show up:

Please enter number of divisions for Weischet plot of Cell #2 (50 max): 20

Please wait while data are analyzed...

57 % done

After calculation, the program will give you the Weischet - Van Holde plot for the sedimentation data. If the boundary has not cleared the meniscus in earlier scans, a slight distortion can appear. Therefore, the analysis can be repeated at this point with a different range of scans included for the analysis.
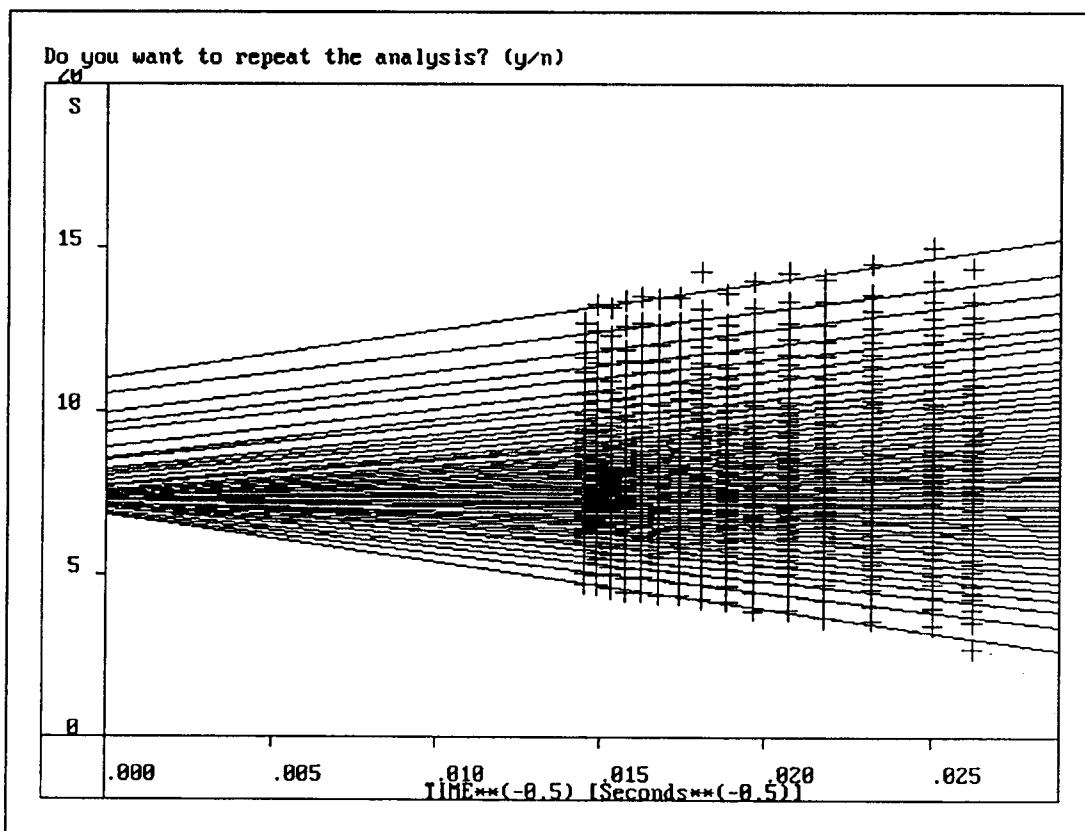
With the mouse you can click around the intersections of the van Holde - Weischet lines with the time=∞ axis to define groups of components. Using the correction data from the hydrodynamic parameter file, an individual sedimentation coefficient for each group is calculated and corrected to water at 20°C. The data is written to an ASCII file of the format

<run identification, cell number,".DAT">

This file can be printed out and it contains all apparent sedimentation coefficients, corrected times, slopes and intercepts for the Weischet plot and the corrected sedimentation coefficients for each group.

```
**********************************************************************
     REPORT OF DATA FROM WEISCHET ANALYSIS FOR RUN 232, CELL #1 :
**********************************************************************
```

Cell #1: CYT. C Oxidase (Triton X) I=0.45

Calculated time of start: 14:39:37 (Correlation coefficient = -.99976)

Calculated sedimentation coefficients:

Scan 1:	Scan 2:	Scan 3:	Scan 4:	Scan 5:	Scan 6:	Scan 7:	Scan 8:
2.710	3.441	3.595	3.669	3.916	3.899	4.167	4.245
3.528	3.962	4.042	4.229	4.370	4.444	4.666	4.702
3.936	4.333	4.551	4.676	4.722	4.806	4.957	4.968
4.262	4.704	4.869	4.899	5.024	5.078	5.165	5.234
4.425	5.001	5.187	5.122	5.275	5.304	5.413	5.462
4.669	5.223	5.314	5.345	5.476	5.530	5.620	5.613
4.914	5.519	5.568	5.568	5.626	5.665	5.827	5.803
5.076	5.667	5.695	5.735	5.826	5.845	5.992	5.992
5.320	5.889	5.885	5.902	6.026	5.980	6.116	6.105
5.483	6.037	6.075	6.068	6.127	6.116	6.240	6.256
5.645	6.185	6.201	6.179	6.276	6.295	6.363	6.331
5.808	6.333	6.391	6.346	6.376	6.385	6.487	6.445
5.970	6.480	6.518	6.457	6.526	6.520	6.610	6.595
6.133	6.628	6.644	6.568	6.676	6.655	6.693	6.708
6.295	6.775	6.770	6.734	6.776	6.745	6.816	6.784
6.376	6.923	6.897	6.845	6.925	6.835	6.898	6.896
6.538	7.070	7.023	6.956	7.025	6.969	7.022	7.009
6.700	7.218	7.149	7.067	7.125	7.059	7.145	7.122
6.781	7.365	7.275	7.177	7.274	7.193	7.227	7.197
6.943	7.512	7.401	7.343	7.373	7.283	7.350	7.310
7.105	7.586	7.528	7.399	7.473	7.372	7.432	7.385
7.186	7.733	7.654	7.564	7.572	7.506	7.555	7.497
7.348	7.880	7.779	7.675	7.672	7.596	7.637	7.610
7.509	8.027	7.905	7.785	7.821	7.685	7.760	7.722
7.590	8.101	8.031	7.896	7.920	7.819	7.842	7.797
7.752	8.247	8.157	8.006	8.019	7.909	7.923	7.909
7.913	8.394	8.283	8.172	8.168	8.043	8.046	8.021
8.075	8.541	8.408	8.282	8.267	8.176	8.169	8.134
8.236	8.688	8.534	8.392	8.366	8.266	8.291	8.246
8.317	8.835	8.723	8.502	8.515	8.355	8.373	8.395
8.478	8.981	8.848	8.667	8.614	8.488	8.495	8.507
8.640	9.128	8.974	8.777	8.762	8.666	8.617	8.619
8.801	9.274	9.099	8.942	8.911	8.800	8.740	8.730
8.962	9.494	9.287	9.052	9.059	8.933	8.862	8.879
9.123	9.640	9.475	9.217	9.207	9.066	9.025	9.028
9.284	9.787	9.600	9.382	9.355	9.200	9.187	9.177
9.445	10.079	9.788	9.601	9.553	9.377	9.350	9.288
9.686	10.225	9.975	9.711	9.700	9.554	9.512	9.474
9.847	10.444	10.163	9.930	9.897	9.687	9.634	9.696
10.088	10.663	10.350	10.149	10.094	9.909	9.836	9.807
10.329	10.882	10.600	10.313	10.291	10.174	10.039	9.992
10.570	11.174	10.849	10.532	10.438	10.351	10.201	10.251
10.810	11.392	11.099	10.805	10.733	10.571	10.484	10.583
11.291	11.829	11.410	11.023	11.125	10.880	10.726	10.804
11.611	12.047	11.721	11.405	11.419	11.144	11.008	11.172
11.931	12.337	12.094	11.677	11.712	11.496	11.330	11.466
12.410	12.845	12.528	12.112	12.201	11.847	11.772	11.979
12.889	13.353	13.086	12.655	12.688	12.460	12.213	12.564
13.367	14.004	13.580	13.359	13.368	13.158	12.653	13.110
14.401	15.014	14.506	14.061	14.241	13.985	13.609	14.271

Scan 9:	Scan 10:	Scan 11:	Scan 12:	Scan 13:	Scan 14:	Scan 15:
4.362	4.362	4.455	4.477	4.579	4.693	4.699
4.750	4.821	4.915	4.908	5.012	4.974	5.038
5.067	5.115	5.159	5.195	5.228	5.305	5.352
5.278	5.311	5.404	5.424	5.471	5.484	5.569
5.523	5.507	5.617	5.624	5.632	5.662	5.738
5.699	5.670	5.769	5.795	5.820	5.865	5.882
5.839	5.865	5.921	5.909	5.955	5.991	6.002
5.944	5.996	6.043	6.052	6.089	6.118	6.122
6.083	6.093	6.164	6.166	6.223	6.219	6.242

A-63

```
6.223    6.255    6.286    6.308    6.303    6.345    6.313
6.363    6.353    6.407    6.393    6.437    6.446    6.409
6.468    6.450    6.498    6.535    6.517    6.547    6.529
6.572    6.580    6.619    6.620    6.624    6.623    6.624
6.677    6.677    6.710    6.706    6.704    6.724    6.719
6.781    6.774    6.800    6.791    6.811    6.799    6.815
6.886    6.871    6.921    6.904    6.917    6.900    6.910
6.990    6.968    7.012    6.989    6.997    7.000    6.982
7.094    7.065    7.132    7.074    7.077    7.076    7.053
7.198    7.194    7.223    7.158    7.157    7.176    7.148
7.268    7.291    7.283    7.243    7.237    7.252    7.219
7.372    7.355    7.404    7.356    7.316    7.327    7.291
7.476    7.452    7.464    7.441    7.423    7.427    7.362
7.545    7.549    7.554    7.525    7.502    7.502    7.433
7.684    7.645    7.644    7.582    7.608    7.577    7.528
7.788    7.710    7.764    7.694    7.688    7.652    7.622
7.892    7.838    7.824    7.807    7.767    7.727    7.693
7.961    7.935    7.944    7.891    7.847    7.802    7.788
8.099    8.031    8.034    8.004    7.926    7.902    7.906
8.203    8.159    8.094    8.088    8.032    8.002    7.977
8.272    8.256    8.214    8.172    8.111    8.077    8.071
8.375    8.352    8.334    8.284    8.217    8.201    8.166
8.513    8.448    8.453    8.396    8.296    8.301    8.260
8.651    8.544    8.573    8.508    8.427    8.450    8.354
8.789    8.672    8.692    8.620    8.507    8.524    8.472
8.892    8.832    8.812    8.732    8.638    8.648    8.589
9.029    8.959    8.931    8.844    8.822    8.772    8.706
9.167    9.087    9.080    8.984    8.927    8.921    8.847
9.338    9.246    9.229    9.123    9.058    9.045    8.941
9.441    9.437    9.377    9.318    9.242    9.218    9.105
9.681    9.628    9.526    9.513    9.399    9.390    9.315
9.921    9.787    9.763    9.735    9.582    9.538    9.478
10.057   9.978    10.000   9.874    9.791    9.784    9.688
10.364   10.199   10.236   10.150   9.999    10.054   9.897
10.603   10.548   10.473   10.399   10.311   10.324   10.245
10.841   10.832   10.826   10.730   10.597   10.617   10.545
11.317   11.178   11.208   11.088   10.933   11.032   10.892
11.655   11.556   11.501   11.527   11.423   11.469   11.191
12.162   12.026   12.028   11.856   11.912   11.833   11.649
12.735   12.651   12.698   12.619   12.296   12.436   12.129
13.473   13.460   13.509   13.378   13.164   13.252   12.698
```

Corrected time for each scan:

```
1. scan: 24.21 minutes
2. scan: 26.54 minutes
3. scan: 30.89 minutes
4. scan: 35.11 minutes
5. scan: 38.91 minutes
6. scan: 43.11 minutes
7. scan: 46.91 minutes
8. scan: 51.11 minutes
9. scan: 55.09 minutes
10. scan: 59.11 minutes
11. scan: 63.09 minutes
12. scan: 67.11 minutes
13. scan: 71.09 minutes
14. scan: 75.11 minutes
15. scan: 79.09 minutes
```

Statistics for line fits:

1. division:
   Slope: -143.9182,   Intercept: 6.8143, Correlation: -.9752

2. division:
   Slope: -119.2425,   Intercept: 6.8190, Correlation: -.9885

3. division:
   Slope: -103.9136,   Intercept: 6.8624, Correlation: -.9841

4. division:
   Slope: -91.9692,   Intercept: 6.8894, Correlation: -.9770

5. division:

```
        Slope:      -85.6455,    Intercept:    6.9906, Correlation:    -.9444

6. division:
        Slope:      -81.0580,    Intercept:    7.0901, Correlation:    -.9495

7. division:
        Slope:      -68.3239,    Intercept:    7.0273, Correlation:    -.9162

8. division:
        Slope:      -65.2322,    Intercept:    7.1166, Correlation:    -.9076

9. division:
        Slope:      -54.1175,    Intercept:    7.0540, Correlation:    -.8802

10. division:
        Slope:      -49.0316,    Intercept:    7.0923, Correlation:    -.8580

11. division:
        Slope:      -44.5377,    Intercept:    7.1305, Correlation:    -.8407

12. division:
        Slope:      -38.5706,    Intercept:    7.1392, Correlation:    -.7991

13. division:
        Slope:      -33.6379,    Intercept:    7.1673, Correlation:    -.7634

14. division:
        Slope:      -27.7473,    Intercept:    7.1676, Correlation:    -.7062

15. division:
        Slope:      -21.4813,    Intercept:    7.1585, Correlation:    -.6245

16. division:
        Slope:      -20.1590,    Intercept:    7.2426, Correlation:    -.5506

17. division:
        Slope:      -14.1897,    Intercept:    7.2391, Correlation:    -.4312

18. division:
        Slope:       -8.2761,    Intercept:    7.2339, Correlation:    -.2720

19. division:
        Slope:       -5.5882,    Intercept:    7.2890, Correlation:    -.1671

20. division:
        Slope:        2.2448,    Intercept:    7.2446, Correlation:     .0687

21. division:
        Slope:        5.1548,    Intercept:    7.2822, Correlation:     .1755

22. division:
        Slope:        8.8285,    Intercept:    7.3200, Correlation:     .2601

23. division:
        Slope:       15.4662,    Intercept:    7.2938, Correlation:     .4369

24. division:
        Slope:       21.3856,    Intercept:    7.2933, Correlation:     .5593

25. division:
        Slope:       21.7048,    Intercept:    7.3826, Correlation:     .5537

26. division:
        Slope:       27.6598,    Intercept:    7.3735, Correlation:     .6593

27. division:
        Slope:       34.7120,    Intercept:    7.3555, Correlation:     .7389

28. division:
        Slope:       38.8639,    Intercept:    7.3931, Correlation:     .7877

29. division:
        Slope:       44.2403,    Intercept:    7.3991, Correlation:     .8323

30. division:
        Slope:       48.5502,    Intercept:    7.4249, Correlation:     .8092
```

31. division:
Slope: 52.8891,   Intercept: 7.4640, Correlation: .8410

32. division:
Slope: 57.7676,   Intercept: 7.4943, Correlation: .8637

33. division:
Slope: 61.2593,   Intercept: 7.5584, Correlation: .8802

34. division:
Slope: 69.0908,   Intercept: 7.5431, Correlation: .8842

35. division:
Slope: 73.3053,   Intercept: 7.6046, Correlation: .8921

36. division:
Slope: 75.2501,   Intercept: 7.7084, Correlation: .9003

37. division:
Slope: 84.2482,   Intercept: 7.7016, Correlation: .8949

38. division:
Slope: 89.7730,   Intercept: 7.7518, Correlation: .9179

39. division:
Slope: 91.9695,   Intercept: 7.8847, Correlation: .9095

40. division:
Slope: 96.2768,   Intercept: 7.9971, Correlation: .9256

41. division:
Slope: 100.7294,   Intercept: 8.1177, Correlation: .9301

42. division:
Slope: 105.4258,   Intercept: 8.2361, Correlation: .9314

43. division:
Slope: 105.1828,   Intercept: 8.4968, Correlation: .9248

44. division:
Slope: 117.5103,   Intercept: 8.5705, Correlation: .9502

45. division:
Slope: 116.4486,   Intercept: 8.8923, Correlation: .9553

46. division:
Slope: 110.0733,   Intercept: 9.3583, Correlation: .9524

47. division:
Slope: 119.1347,   Intercept: 9.6081, Correlation: .9471

48. division:
Slope: 127.1141,   Intercept: 9.9471, Correlation: .9411

49. division:
Slope: 127.7808,   Intercept: 10.5206, Correlation: .9089

50. division:
Slope: 147.9636,   Intercept: 10.9950, Correlation: .8982

V-BAR for your Sample: .740
Density of your Buffer at 29.81 °C: .995
Viscosity of your Buffer at 29.81 °C: .815
Rotor Speed: 40000.000

Calculated sedimentation coefficient from 1. group: 10.0858
Sedimentation coefficient corrected for water at 20°C from 1. group: 8.08958

Calculated sedimentation coefficient from 2. group: 7.42907
Sedimentation coefficient corrected for water at 20°C from 2. group: 5.95865

The next two graphics will provide the uncorrected and corrected distribution plots for the van Holde - Weischet analysis:

Integral Distribution of Cell #1

And the corrected distribution plot:

Integral Distribution of Cell #1
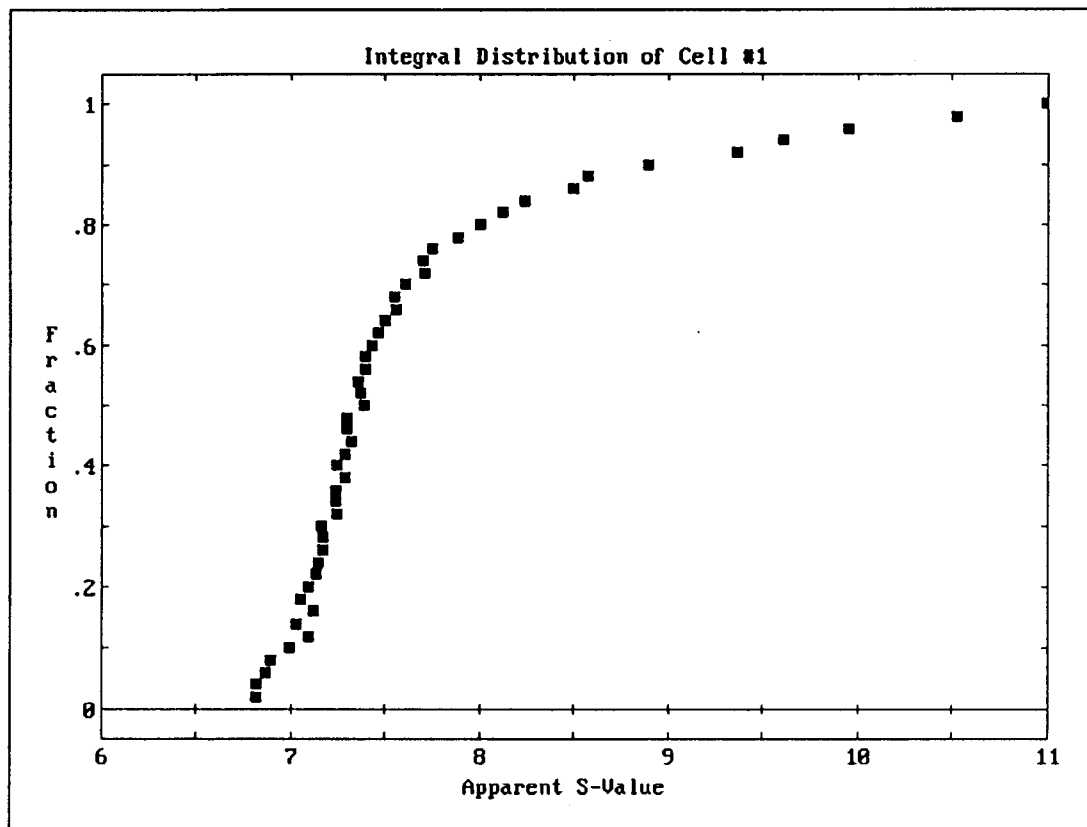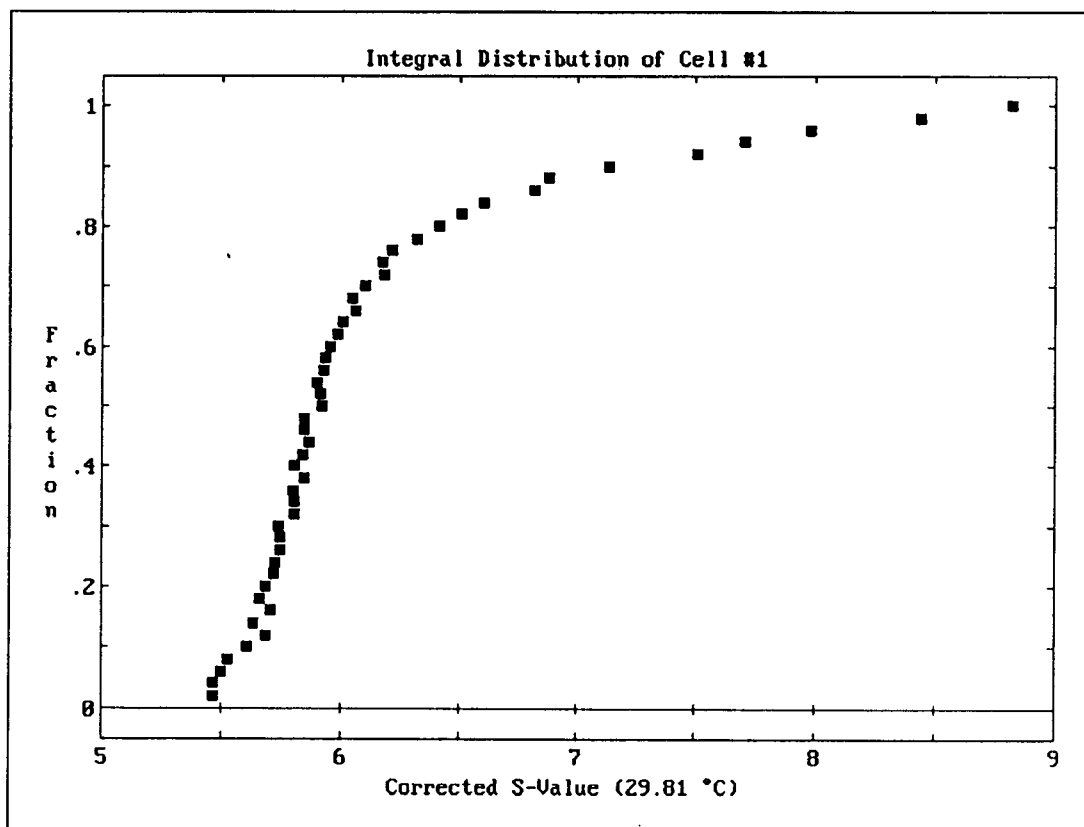
Following the distribution plots the option to print the data is available. The data can either be sent directly to the printer or the output can be sent to a file for later incorporation into a word processor file. Choose between "f" (for file) and "p" (for printer). The following menu will appear:

```
┌─────────────────────────────────────────────────────────────┐
│          van Holde - Weischet Analysis Printing Program:     │
│                                                               │
├─────────────────────────────────────────────────────────────┤
│  ┌─────────────────────────────────────────────────────┐    │
│  │                    PRINT MENU:                       │    │
│  │                                                       │    │
│  │  F1: Print van Holde - Weischet Analysis Data Report │    │
│  │  F2: Print van Holde - Weischet Analysis Plot        │    │
│  │  F3: Print Apparent S-Value Distribution             │    │
│  │  F4: Print Corrected S-Value Distribution            │    │
│  │  F5: Print Edited Scans                               │    │
│  │  F6: Exit Program                                     │    │
│  └─────────────────────────────────────────────────────┘    │
│                                                               │
│                                                               │
└─────────────────────────────────────────────────────────────┘
```

The printing of graphic screendumps requires prior loading of a suitable graphics screen driver program such as GRAPHICS.COM provided by MS-DOS or a more flexible program like PIZAZZ PLUS. Just choose the menu item and obtain the desired printout. For high resolution printouts, the data is saved in a metafile, which can be imported into programs like SUPERCALC, SIGMAPLOT, LOTUS 1-2-3, QUATTRO PRO etc. For more information please consult the section on metafiles.

SECOND MOMENT ANALYSIS:

The second moment analysis will calculate the weighted average sedimentation coefficient for all components in the cell from edited sedimentation velocity datafiles. A separate calculation is performed for each scan of each cell. The second moment points are calculated by integration according to following formula:

$$\left|\overline{R}\right|^2 = \frac{\int R^2 * \frac{\delta C}{\delta R} \, dR}{\int \frac{\delta C}{\delta R} \, dR}$$

To start the second moment analysis select "Second Moment Analysis" from the Analysis Program Menu and choose the run identification, number of cells and scans to be included in the analysis. Determine the proper hydrodynamic parameter corrections for sedimentation coefficient correction to water at 20° C. The edited scans of each cell will appear on the screen, and you have to click at the point which will be the limit of the integration, usually at the bottom of the cell:

Please use the mouse to define the left limit for the integration:
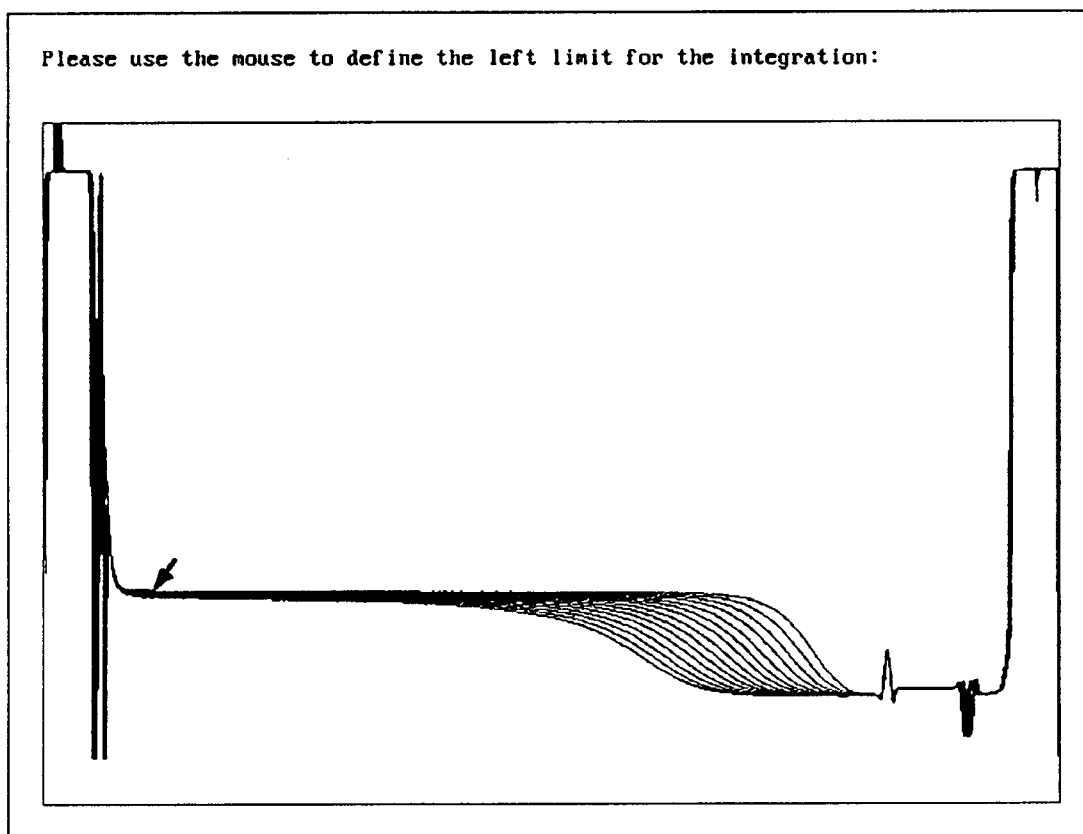
You have the option to send the output of the second moment program to the printer. Whether you choose to print the data or not, the results from the second moment analysis will automatically written to the file <run identification, cell number,".DAT"> or appended to it, if it exists already from the van Holde - Weischet analysis.

The calculated midpoints are compared to the estimated midpoints and the differences are listed. The estimated midpoints and calculated second moment points are displayed together with the temperature/buffer corrected sedimentation coefficients, and the differences are listed:

```
V-BAR for your Sample:                              .740
Density of your Buffer at 29.81 °C:                 .995
Viscosity of your Buffer at 29.81 °C:               .815
Rotor Speed:                                   40000.000
```

Second Moment Analysis for Run ID: 232, Cell #1
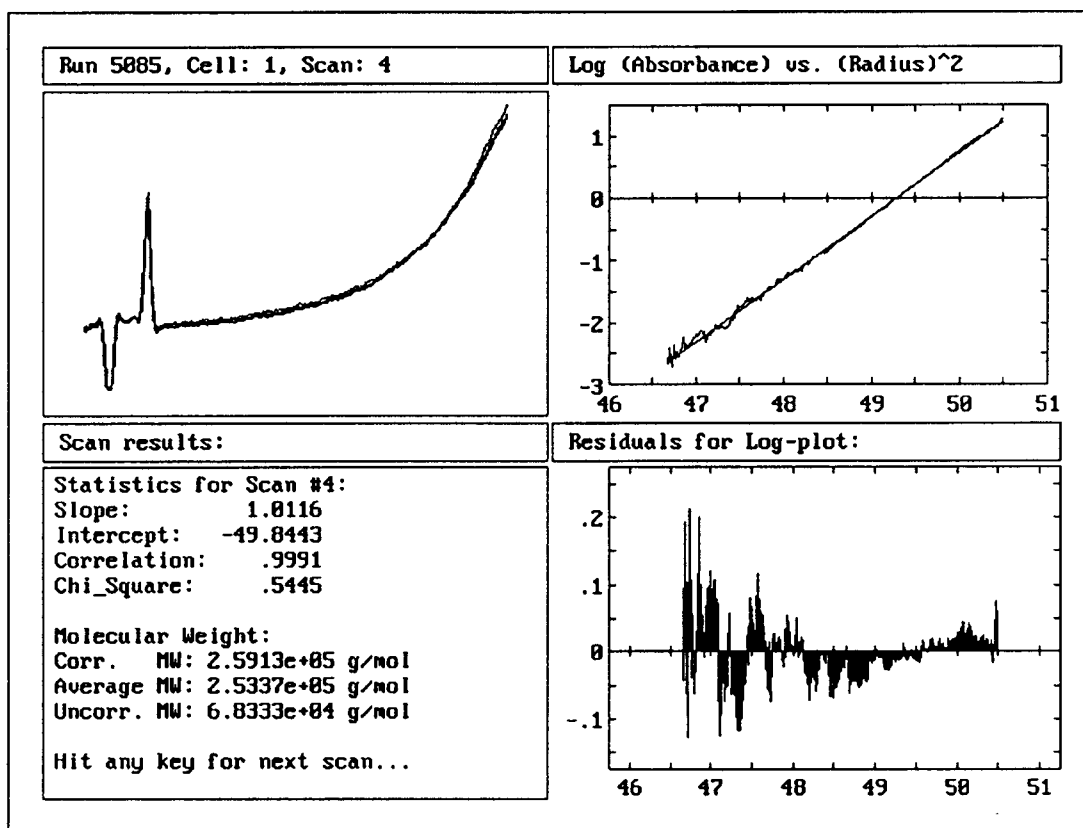
Cell #1: CYT. C Oxidase (Triton X) I=0.45

Scan:	Estm. Mpoint:	Calc. Mpoint:	Difference:	Calc. S-Val.:	Corr. S-Val.:
1	6.083 cm	6.107 cm	- .024 cm	8.898	7.137
2	6.106 cm	6.134 cm	- .029 cm	9.674	7.759
3	6.126 cm	6.156 cm	- .030 cm	9.424	7.559
4	6.146 cm	6.178 cm	- .032 cm	9.240	7.411
5	6.167 cm	6.200 cm	- .034 cm	9.232	7.405
6	6.185 cm	6.222 cm	- .037 cm	9.118	7.313
7	6.205 cm	6.242 cm	- .038 cm	9.029	7.242
8	6.225 cm	6.267 cm	- .042 cm	9.015	7.231
9	6.245 cm	6.289 cm	- .044 cm	8.970	7.195
10	6.264 cm	6.306 cm	- .042 cm	8.802	7.060
11	6.284 cm	6.327 cm	- .043 cm	8.746	7.015
12	6.306 cm	6.356 cm	- .050 cm	8.862	7.108
13	6.326 cm	6.376 cm	- .050 cm	8.780	7.043
14	6.345 cm	6.396 cm	- .050 cm	8.707	6.984
15	6.363 cm	6.415 cm	- .051 cm	8.626	6.919

The average values for Cell #1 are: - .040 cm   9.008   7.225

SEDIMENTATION EQUILIBRIUM ANALYSIS:

The sedimentation equilibrium analysis performs a single component logarithmic fit on the equilibrium data. The baseline is automatically optimized by minimizing the chi-square residuals from the logarithmic fit. After selecting the data, you have to select the hydrodynamic parameters to make accurate temperature and buffer corrections. The equilibrium data analysis presents three graphs: Absorbance vs. radius, log(absorbance) vs. (radius)^2, the residuals from the log-plot and the calculated results and statistics for each scan. The purpose of the residuals plot is to inform the user if the single component fit is an acceptable method for analyzing the data. If the residuals are spread about 0 randomly, the single component fit provides adequate information. If, however, the residuals display a 'run' pattern, i.e., many subsequent points have the same sign, another method of analysis should be chosen.

The statistics of the run provide the slope of the fitted line, the y-intercept and the correlation coefficient and the chi-square value of the fitted log-plot of the scan. The molecular weight is given with and without buoyancy correction and average molecular weight of all scans included in the analysis is calculated. The averaging of several scans of the same sample is recommended to overcome eventual electronic noise bias.

The edited data plot on the upper left is intended to determine the status of the run. If the sample has essentially approached equilibrium, subsequent scans spaced a few hours apart should overlap exactly. If a difference (a thickening of the absorbance trace) is observed, the early scans are not at equilibrium. The fact that scans can be overlaid should help the user to determine if the sample has effectively reached equilibrium. Also, if equilibrium has been reached, the calculated molecular weight should not change from scan to scan.

```
┌─────────────────────────────────────┬─────────────────────────────────────┐
│ Run 5085, Cell: 1, Scan: 4          │ Log (Absorbance) vs. (Radius)^2     │
├─────────────────────────────────────┼─────────────────────────────────────┤
```



```
┌─────────────────────────────────────┐
│ Scan results:                       │   Residuals for Log-plot:
├─────────────────────────────────────┤
│ Statistics for Scan #4:             │
│ Slope:          1.0116              │
│ Intercept:     -49.8443             │
│ Correlation:     .9991              │
│ Chi_Square:      .5445              │
│                                     │
│ Molecular Weight:                   │
│ Corr.   MW: 2.5913e+05 g/mol        │
│ Average MW: 2.5337e+05 g/mol        │
│ Uncorr. MW: 6.8333e+04 g/mol        │
│                                     │
│ Hit any key for next scan...        │
└─────────────────────────────────────┘
```

The equilibrium data is saved to a file called <run number, cell number, ".RES">, and it contains the results for each scan in the cell:

```
**********************************************************************
       EQUILIBRIUM ANALYSIS FOR RUN 5085, CELL #1 :
**********************************************************************

Cell #1: Sed. Eq. run for RsaI tetramer(scan taken 10:20 4/25)

V-BAR for your Sample:                        .740
Density of your Buffer at 12.13 °C:           .995
Viscosity of your Buffer at 12.13 °C:         .815
Rotor Speed:                              8000.000


Statistics for Scan # 1 :

Slope:         .9503
Intercept: -  46.8301
Correlation:   .9992
Chi_Square:    .3967


Molecular Weight:

Corr.   MW: 2.4344e+05 g/mol
Average MW: 2.4344e+05 g/mol
Uncorr. MW: 6.4196e+04 g/mol


Statistics for Scan # 2 :
```

```
Slope:           .9877
Intercept: -   48.7022
Correlation:     .9993
Chi_Square:      .3607
```

Molecular Weight:

```
Corr.   MW: 2.5303e+05 g/mol
Average MW: 2.4823e+05 g/mol
Uncorr. MW: 6.6723e+04 g/mol
```

Statistics for Scan # 3 :

```
Slope:          1.0067
Intercept: -   49.6308
Correlation:     .9993
Chi_Square:      .3746
```

Molecular Weight:

```
Corr.   MW: 2.5790e+05 g/mol
Average MW: 2.5146e+05 g/mol
Uncorr. MW: 6.8008e+04 g/mol
```

Statistics for Scan # 4 :

```
Slope:          1.0116
Intercept: -   49.8443
Correlation:     .9991
Chi_Square:      .5445
```

Molecular Weight:

```
Corr.   MW: 2.5913e+05 g/mol
Average MW: 2.5337e+05 g/mol
Uncorr. MW: 6.8333e+04 g/mol
```

META FILE GENERATION:

The ANALYSIS PROGRAM provides metafiles for import into graphing programs such as SIGMAPLOT, LOTUS 1-2-3, SUPERCALC or QUATTRO-PRO. Metafiles are ASCII formatted files, which contain comma delimited data providing all necessary information for high resolution plotting. Metafiles are generated for the van Holde - Weischet analysis plot, the van Holde - Weischet distribution plots and the equilibrium data plots.

1. Metafile for van Holde - Weischet analysis:

The name for this file is <run number, cell number, ".MET">, and the data contained is as follows: The first column contains the inverse square - root of time for each scan, followed by the apparent sedimentation data written into the same column. For column separation use the appropriate column seperation program within the graphic program. The number of columns should be the number of integral divisions for the van Holde - Weischet plot plus 1 column for the inverse square-root of the time. The number of elements in each column is the number of scans included in the analysis.

2. Metafile for the van Holde - Weischet distribution plots:

The name for this file is <run number, cell number, ".DIS">, and the data contained is as follows: There are three comma-delimited columns, the first column contains the uncorrected distribution S-values, the second column contains the buffer and temperature corrected distribution S-values and the last column contains the fractional values of each S-value. There will be as many elements in each column as there are divisions in the analysis.

3. Metafile for the equilibrium analysis:

The name for the equilibrium data file is <run number, cell number, "." scan number, "T">, and the data contained is as follows: There are three comma-delimited columns, containing the (radius)^2, log-data and the residuals.

As an alternative, the data can be printed with the printing program, using a low resolution screen dump routine. A graphics screen driver has to be loaded into memory before loading the ANALYSIS PROGRAM. A suitable driver is GRAPHICS.COM or a program such as PIZAZZ PLUS, which allows a greater flexibility.
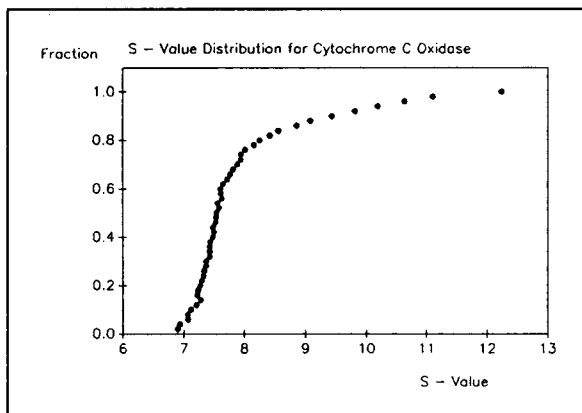
Some examples for higher resolution obtained from metafiles follow:
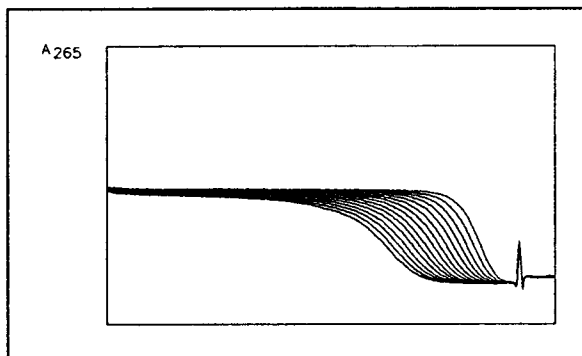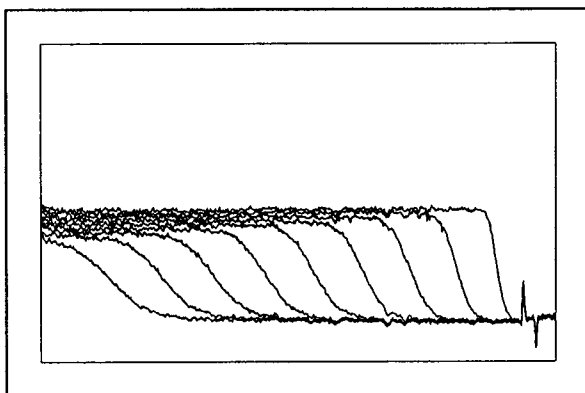
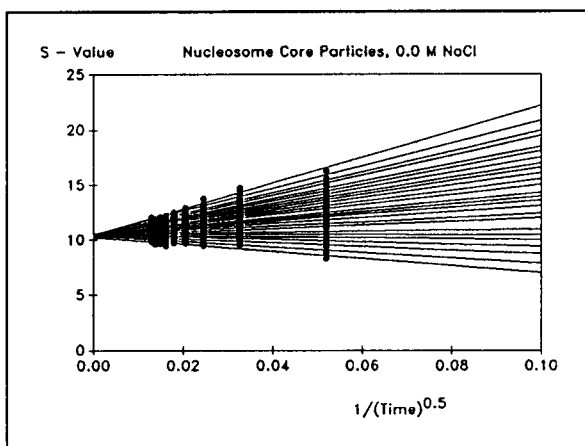**Cytochrome C Oxidase**
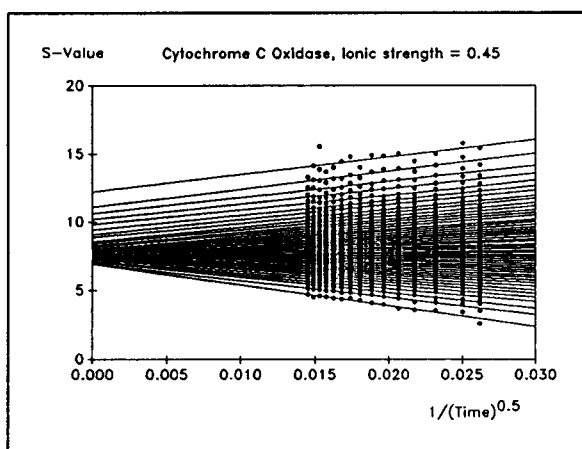
van Holde - Weischet Plot
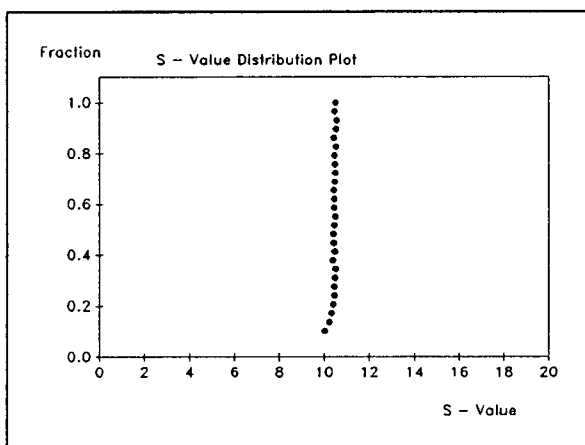
Integral Distribution Plot

van Holde - Weischet analysis metafile output.
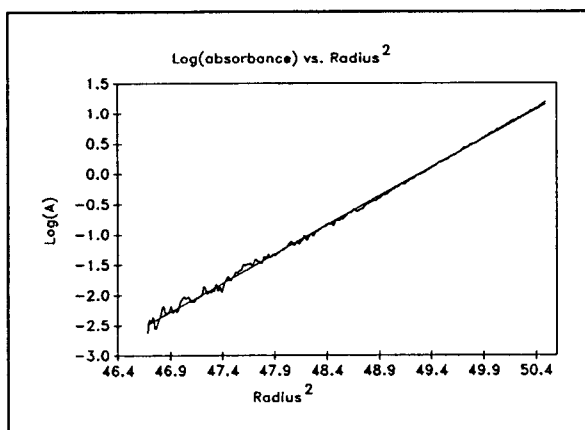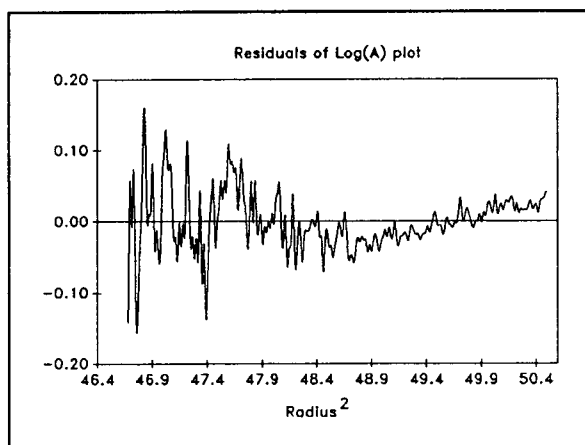
Nucleosome Core Particles

van Holde - Weischet Plot

Integral Distribution Plot

van Holde - Weischet analysis metafile output.

Log(A) vs. Radius² Plot

Residuals of Log(A) Plot

Equilibrium analysis metafile output.

(All plots have been prepared with SIGMAPLOT)

MODEL-E INTERFACE HARDWARE INSTALLATION INSTRUCTIONS:

Tools required:

Screwdriver
Soldering iron
Epoxy glue

Carefully unpack the interface, making sure not ot bend/break any wires. Rout the 4-lead cable through the back of the Model-E into the scanner compartment. Unscrew the wiremesh cover and remove. Locate Terminal Bank 1 and 2 (see schematic) in the right back of scanner unit. Remove the terminal covers and determine signal terminal (should be terminal 19 on Bank 2) and ground terminal (should be terminal 17 on Bank 2).

ATTENTION:

Verify proper terminals by measuring voltage during run (Careful, do not touch hot leads!) before fastening the signal wire and the ground wire. The voltage should be approximately between 0 and -10 V DC during the run.

Fasten signal (red wire) to terminal 19 on Bank 2 and ground wire (black) to terminal 17 on Bank 2. Next, find a suitable spot for the trigger relais and glue it with epoxy to the frame. Make sure that no wires will touch the cover or interfere with other components of the scanner. Solder the red wire from the relais labelled S602 E-5 to the Multi-Switch S-602 on Bank E-5. To determine proper position, consult the schematic. Connect the white wire labelled TB1-12 from the relais to Bank 1 on terminal 12. Plug the 5-Pin DIN connector into the back of the interface box and connect the DB25 socket into the Metrabyte CROM-1 At port. Make sure that interrupts, triggers, relais switches and base address is set properly in the "Program Configuration" function from the UTILITY PROGRAM.

WIRING SCHEMATIC FOR MODEL-E INTERFACE:
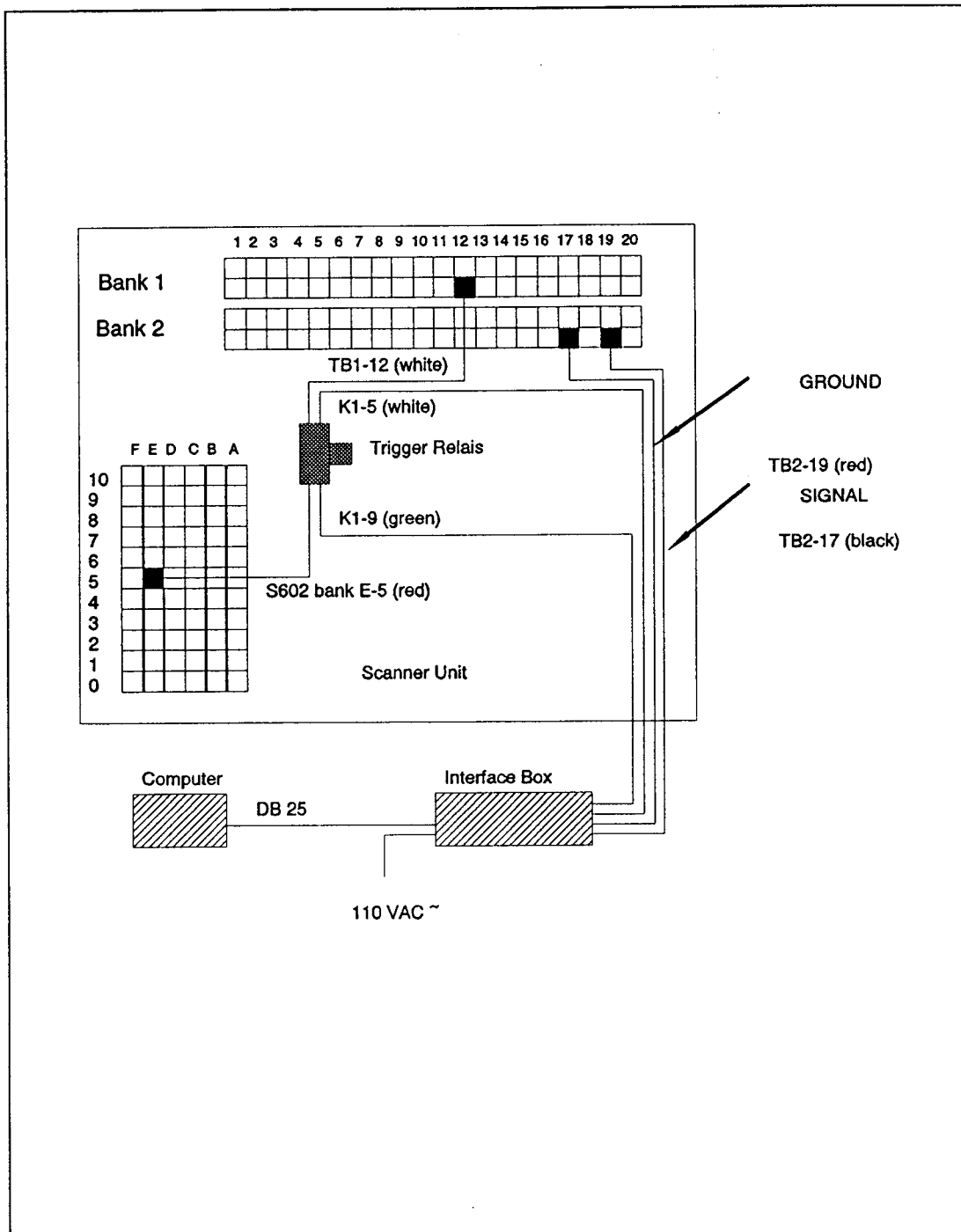
APPENDIX B - PROGRAM BLINE.FOR

```fortran
      PROGRAM BLINE.FOR
      COMMON /BASE1/ X(1024)
      COMMON /BASE2/ Y(1024)
      COMMON /BASE3/ N
      COMMON /BASE4/ YLOG(1024)

      OPEN(UNIT=1,FILE='TR_IN.DAT',FORM='FORMATTED',STATUS='OLD')
      READ (1,100) N
100   FORMAT (1I4)
      CLOSE (1)

C READ THE DATA: X=REAL RADIUS, Y=ABSORBANCE, N=NUMBER OF POINTS,
C BLINE=ESTIMATED BASELINE

      CALL READ_DATA(X,Y,N,BLINE)

C LIMITS AMONG WHICH TO SEARCH FOR OPTIMIAL BASELINE (ESTIMATE:
C BLINE)

      START_B=BLINE-1.
      STOP_B=BLINE+1.

C TOLERA IS ACCURACY TO WHICH BASELINE IS OPTIMIZED

      TOLERA=0.001

C CALC_BASE IS THE OPTIMIZED BASELINE VALUE FOR THE LOG(A) PLOT

      RES=GOLDEN(START_B,BLINE,STOP_B,TOLERA,CALC_BASE)

C SUBTRACT BASELINE FROM ABSORBANCE DATA AND STORE IN YLOG()

      DO 150 J=1,N
         YLOG(J)=LOG(ABS(Y(J)-CALC_BASE))
150   CONTINUE
      CALL WRITE_DATA(CALC_BASE,RES)
      END

********************************************************************

      FUNCTION RESIDUALS (BX)
      COMMON /BASE1/ X(1024)
      COMMON /BASE2/ Y(1024)
      COMMON /BASE3/ N
      COMMON /BASE4/ YLOG(1024)

      SUM=0.
      DO 100 J=1,N
```

```
            YLOG(J)=LOG(ABS(Y(J)-BX))
100     CONTINUE
        MWT=0
        CALL FIT(X,YLOG,N,SIG,MWT,A,B,SIGA,SIGB,CHI2,Q)
        RESIDUALS=CHI2
        RETURN
        END


*************************************************************

        SUBROUTINE READ_DATA(X,Y,N,BLINE)
        DIMENSION X(N),Y(N)

        OPEN(UNIT=1,FILE='TR_IN.DAT',FORM='FORMATTED',STATUS='OLD')
        READ (1,100) N
        READ (1,200) BLINE
100     FORMAT (1I4)
200     FORMAT (1F10.6)
        DO 300 J = 1, N
            READ (1,400) X(J), Y(J)
300     CONTINUE
400     FORMAT (1F10.6,1X,1F10.6)
        CLOSE (1)
        RETURN
        END


*************************************************************

        SUBROUTINE WRITE_DATA(CALC_BASE,RES)

        COMMON /BASE2/ Y(1024)
        COMMON /BASE3/ N
        COMMON /BASE4/ YLOG(1024)

        OPEN (UNIT=1,FILE='TR_OUT.DAT',FORM='FORMATTED',STATUS='OLD')
        WRITE (1,100) CALC_BASE
        WRITE (1,100) RES

100     FORMAT (1E16.6)
        DO 300 J = 2, N
            WRITE (1,400) YLOG(J)
300     CONTINUE

C Y=ABSORBANCE
C YLOGUS=UNSMOOTHED LOG(A)

400     FORMAT (1E13.6)
        CLOSE (1)
        RETURN
        END
```

```
*****************************************************************
        FUNCTION GOLDEN(AX,BX,CX,TOL,XMIN)
        PARAMETER (R=.61803399,C=.38196602)
        X0=AX
        X3=CX
        IF(ABS(CX-BX).GT.ABS(BX-AX))THEN
           X1=BX
           X2=BX+C*(CX-BX)
        ELSE
           X2=BX
           X1=BX-C*(BX-AX)
        ENDIF
        F1=RESIDUALS(X1)
        F2=RESIDUALS(X2)
1       IF(ABS(X3-X0).GT.TOL*(ABS(X1)+ABS(X2)))THEN
           IF(F2.LT.F1)THEN
              X0=X1
              X1=X2
              X2=R*X1+C*X3
              F0=F1
              F1=F2
              F2=RESIDUALS(X2)
           ELSE
              X3=X2
              X2=X1
              X1=R*X2+C*X0
              F3=F2
              F2=F1
              F1=RESIDUALS(X1)
           ENDIF
        GOTO 1
        ENDIF
        IF(F1.LT.F2)THEN
           GOLDEN=F1
           XMIN=X1
        ELSE
           GOLDEN=F2
           XMIN=X2
        ENDIF
        RETURN
        END


*****************************************************************
        SUBROUTINE FIT(X,Y,NDATA,SIG,MWT,A,B,SIGA,SIGB,CHI2,Q)
        DIMENSION X(NDATA),Y(NDATA),SIG(NDATA)
        SX=0.
        SY=0.
        ST2=0.
        B=0.
```

```
        IF(MWT.NE.0) THEN
          SS=0.
          DO 11 I=1,NDATA
            WT=1./(SIG(I)**2)
            SS=SS+WT
            SX=SX+X(I)*WT
            SY=SY+Y(I)*WT
11        CONTINUE
        ELSE
          DO 12 I=1,NDATA
            SX=SX+X(I)
            SY=SY+Y(I)
12        CONTINUE
          SS=FLOAT(NDATA)
        ENDIF
        SXOSS=SX/SS
        IF(MWT.NE.0) THEN
          DO 13 I=1,NDATA
            T=(X(I)-SXOSS)/SIG(I)
            ST2=ST2+T*T
            B=B+T*Y(I)/SIG(I)
13        CONTINUE
        ELSE
          DO 14 I=1,NDATA
            T=X(I)-SXOSS
            ST2=ST2+T*T
            B=B+T*Y(I)
14        CONTINUE
        ENDIF
        B=B/ST2
        A=(SY-SX*B)/SS
        SIGA=SQRT((1.+SX*SX/(SS*ST2))/SS)
        SIGB=SQRT(1./ST2)
        CHI2=0.
        IF(MWT.EQ.0) THEN
          DO 15 I=1,NDATA
            CHI2=CHI2+(Y(I)-A-B*X(I))**2
15        CONTINUE
          Q=1.
          SIGDAT=SQRT(CHI2/(NDATA-2))
          SIGA=SIGA*SIGDAT
          SIGB=SIGB*SIGDAT
        ELSE
          DO 16 I=1,NDATA
            CHI2=CHI2+((Y(I)-A-B*X(I))/SIG(I))**2
16        CONTINUE
          Q=GAMMQ(0.5*(NDATA-2),0.5*CHI2)
        ENDIF
        RETURN
        END
```

B-4

```
       FUNCTION GAMMQ(A,X)
       IF(X.LT.0..OR.A.LE.0.)PAUSE
       IF(X.LT.A+1.)THEN
          CALL GSER(GAMSER,A,X,GLN)
          GAMMQ=1.-GAMSER
       ELSE
          CALL GCF(GAMMCF,A,X,GLN)
          GAMMQ=GAMMCF
       ENDIF
       RETURN
       END

       SUBROUTINE GCF(GAMMCF,A,X,GLN)
       PARAMETER (ITMAX=100,EPS=3.E-7)
       GLN=GAMMLN(A)
       GOLD=0.
       A0=1.
       A1=X
       B0=0.
       B1=1.
       FAC=1.
       DO 11 N=1,ITMAX
          AN=FLOAT(N)
          ANA=AN-A
          A0=(A1+A0*ANA)*FAC
          B0=(B1+B0*ANA)*FAC
          ANF=AN*FAC
          A1=X*A0+ANF*A1
          B1=X*B0+ANF*B1
          IF(A1.NE.0.)THEN
             FAC=1./A1
             G=B1*FAC
             IF(ABS((G-GOLD)/G).LT.EPS)GO TO 1
             GOLD=G
          ENDIF
11     CONTINUE
       PAUSE 'A TOO LARGE, ITMAX TOO SMALL'
1      GAMMCF=EXP(-X+A*ALOG(X)-GLN)*G
       RETURN
       END

       SUBROUTINE GSER(GAMSER,A,X,GLN)
       PARAMETER (ITMAX=100,EPS=3.E-7)
       GLN=GAMMLN(A)
       IF(X.LE.0.)THEN
          IF(X.LT.0.)PAUSE
          GAMSER=0.
          RETURN
       ENDIF
       AP=A
       SUM=1./A
```

```
       DEL=SUM
       DO 11 N=1,ITMAX
         AP=AP+1.
         DEL=DEL*X/AP
         SUM=SUM+DEL
         IF(ABS(DEL).LT.ABS(SUM)*EPS)GO TO 1
11     CONTINUE
       PAUSE 'A TOO LARGE, ITMAX TOO SMALL'
1      GAMSER=SUM*EXP(-X+A*LOG(X)-GLN)
       RETURN
       END

       FUNCTION GAMMLN(XX)
       REAL*8 COF(6),STP,HALF,ONE,FPF,X,TMP,SER
       DATA COF,STP/76.18009173D0,-86.50532033D0,24.01409822D0,
     *  -1.231739516D0,.120858003D-2,-.536382D-5,2.50662827465D0/
       DATA HALF,ONE,FPF/0.5D0,1.0D0,5.5D0/
       X=XX-ONE
       TMP=X+FPF
       TMP=(X+HALF)*LOG(TMP)-TMP
       SER=ONE
       DO 11 J=1,6
         X=X+ONE
         SER=SER+COF(J)/X
11     CONTINUE
       GAMMLN=TMP+LOG(STP*SER)
       RETURN
       END
```

APPENDIX C - PROGRAM LEV-MAR.FOR

```
C***********************************************************
C LANGLEY FORD LASERSCAN ANALYSIS PROGRAM LEV-MAR.FOR VER. 1.10 *
C
C***********************************************************

       PROGRAM LEVENBERG_MARQUARDT

       REAL A(10),SIGMAY(1024),FLAMDA
       REAL SIGMAA(10),YFIT(1024),TOL,DIFF,CHISQR,FX
       REAL X(1024),Y(1024)
       INTEGER NPTS,MODE,NTERMS,FUNC_TYPE,ITERA

C INITIALIZE PARAMETER DEFINITIONS:

       FLAMDA=0.0001

C READ THE DATA: X=REAL TIME, Y=INTENSITY, NDATA=NUMBER OF POINTS,
C A(1-10)=PARAMETERS TO BE FITTED (ALSO SERVING AS INITIAL GUESSES)

C NTERMS=NUMBER OF PARAMETERS TO BE FITTED, MODE=TYPE OF WEIGHTING:

C 1: WEIGHT=1/Y
C 2: WEIGHT=1
C 3: WEIGHT=1/STANDARD DEVIATION(Y)
C 4: WEIGHT=1/DATAPOINT INDEX
C 5: WEIGHT=LINEARLY DECREASING
C FUNC_TYPE=SELECTS WHICH FUNCTION IS FITTED:
C 1: Y=A1*X+A2                          (LINEAR FIT)
C 2: Y=A1*EXP(A2*X)+A3                  (SINGLE EXPONENTIAL)
C 3: Y=A1*EXP(A2*X)+A3*EXP(A4*X)+A5     (DOUBLE EXPONENTIAL)
C 4: Y=A1*EXP(A2*X)+A3*EXP(A4*X)+A5*EXP(A(6)*X)+A(7)
C                                       (TRIPLE EXPONENTIAL)
C 5: Y=C*K1**3*(8-7*EXP(K2/X))

       CALL READ_DATA(X,Y,NPTS,A,NTERMS,MODE,TOL,FUNC_TYPE)

C INITIALIZE STANDARD DEVIATIONS TO 1.0

       DO 20 J=1,NPTS
          SIGMAY(J)=1.0
20     CONTINUE

C MAIN LOOP:

       ITERA=0
       FX=1.0E+10
       DIFF=1.0E+10
       WRITE (*,40)
```

```
        WRITE (*,40)
        TEMP=1.0E+10
30      IF (ABS(FX) .GT. TOL) THEN
           ITERA=ITERA+1

           CALL CURVFIT (X,Y,SIGMAY,NPTS,NTERMS,MODE,
     +     A,SIGMAA,FLAMDA,YFIT,CHISQR,FUNC_TYPE)
           DIFF=TEMP-CHISQR
           FX=DIFF/(CHISQR/(NPTS-NTERMS-1))
           WRITE (*,50) ITERA,FX,CHISQR
           TEMP=CHISQR
           GOTO 30
        ENDIF
        CALL WRITE_DATA(SIGMAA,A,NTERMS,FX,CHISQR,ITERA)
        WRITE(*,60) '
     +
     +                                                      '
40      FORMAT ('0')
50      FORMAT ('+','ITERATION:',1I5,', F-TEST:',1E15.7,
     +', X-SQUARE:',1E15.7)
60      FORMAT ('+',A79)
        END


C****************************************************************

        SUBROUTINE READ_DATA(X,Y,NPTS,A,NTERMS,MODE,TOL,FUNC_TYPE)

        REAL A(10),TOL
        REAL X(1024),Y(1024)
        INTEGER NPTS,NTERMS,MODE,FUNC_TYPE

        OPEN(UNIT=1,FILE='INPUT.DAT',FORM='FORMATTED',STATUS='OLD')

C READ INTEGERS:

        READ (1,30) NPTS
        READ (1,30) NTERMS
        READ (1,30) MODE
        READ (1,30) FUNC_TYPE
C          PRINT*,'NPTS: ', NPTS
C          PRINT*,'NTERMS: ', NTERMS
C          PRINT*,'MODE: ', MODE
C          PRINT*,'FUNC_TYPE: ', FUNC_TYPE

C READ REALS:

        READ (1,40) TOL
        DO 10 J=1,NTERMS
           READ (1,40) A(J)
10      CONTINUE
```

```
C READ X AND Y COORDINATE DATA:

        DO 20 J=1,NPTS
           READ (1,50) X(J), Y(J)
C          PRINT*,X(J),Y(J)
20         CONTINUE

C FORMAT STATEMENTS:

30      FORMAT (1I4)
40      FORMAT (1F14.7)
50      FORMAT (1F14.7,1X,1F14.7)
        CLOSE (1)
        RETURN
        END


C*************************************************************

        SUBROUTINE WRITE_DATA(SIGMAA,A,NTERMS,FX,CHISQR,ITERA)

        REAL SIGMAA(10),A(10),FX,CHISQR
        INTEGER NTERMS,I,ITERA

        OPEN (UNIT=1,FILE='OUTPUT.DAT',FORM='FORMATTED')
        WRITE (1,30) FX
        WRITE (1,30) CHISQR
        WRITE (1,40) ITERA
        DO 10 I=1,NTERMS
           WRITE (1,30) A(I)
10      CONTINUE
        DO 20 I=1,NTERMS
           WRITE (1,30) SIGMAA(I)
20      CONTINUE
30      FORMAT (1E16.6)
40      FORMAT (1I16)
        CLOSE (1)
        RETURN
        END


C*************************************************************

        SUBROUTINE CURVFIT (X,Y,SIGMAY,NPTS,NTERMS,MODE,
      + A,SIGMAA,FLAMDA,YFIT,CHISQR,FUNC_TYPE)

        DOUBLE PRECISION ARRAY
        REAL SIGMAY(1024),YFIT(1024),A(10),BETA(10)
        REAL DERIV(10),SIGMAA(10),B(10)
        REAL CHISQR,CHISQ1,ALPHA(10,10),FLAMDA
        REAL X(1024),Y(1024),WEIGHT(1024)
        INTEGER NPTS,NTERMS,NFREE,FUNC_TYPE
        DIMENSION ARRAY(10,10)
```

C-3

```
       NFREE=NPTS-NTERMS
       IF(NFREE) 10,10,20
10     CHISQR=0.
       GOTO 9999
20     CALL WEIGHT_ASSIGN(WEIGHT,Y,MODE,SIGMAY,NPTS)
       DO 110 J=1,NTERMS
          BETA(J)=0.0
          DO 100 K=1,J
             ALPHA(J,K)=0.0
100       CONTINUE
110    CONTINUE
       DO 140 I=1,NPTS
          CALL FDERIV(X,I,A,DERIV,FUNC_TYPE)
          DO 130 J=1,NTERMS
             BETA(J)=BETA(J)+WEIGHT(I)*(Y(I)-
     +       FUNCTN(X,I,A,FUNC_TYPE))*DERIV(J)
             DO 120 K=1,J
                ALPHA(J,K)=ALPHA(J,K)+WEIGHT(I)*DERIV(J)*DERIV(K)
120          CONTINUE
130       CONTINUE
140    CONTINUE
       DO 160 J=1,NTERMS
          DO 150 K=1,J
             ALPHA(K,J)=ALPHA(J,K)
150       CONTINUE
160    CONTINUE
       DO 170 I=1,NPTS
          YFIT(I)=FUNCTN(X,I,A,FUNC_TYPE)
170    CONTINUE
       CHISQ1=FCHISQ(Y,SIGMAY,NPTS,NFREE,MODE,YFIT)
175    DO 190 J=1,NTERMS
          DO 180 K=1,NTERMS
             ARRAY(J,K)=ALPHA(J,K)/SQRT(ALPHA(J,J)*ALPHA(K,K))
180       CONTINUE
          ARRAY(J,J)=1.0+FLAMDA
190    CONTINUE
       CALL MATINV(ARRAY,NTERMS,DET)
       DO 210 J=1,NTERMS
          B(J)=A(J)
          DO 200 K=1,NTERMS
             B(J)=B(J)+BETA(K)*ARRAY(J,K)/SQRT(ALPHA(J,J)*ALPHA(K,K))
200       CONTINUE
210    CONTINUE
       DO 220 I=1,NPTS
          YFIT(I)=FUNCTN(X,I,B,FUNC_TYPE)
220    CONTINUE
       CHISQR=FCHISQ(Y,SIGMAY,NPTS,NFREE,MODE,YFIT)
       IF (CHISQ1-CHISQR) 230,240,240
230    FLAMDA=10*FLAMDA
       GOTO 175
240    DO 250 J=1,NTERMS
```

```
              A(J)=B(J)
              SIGMAA(J)=SQRT(ARRAY(J,J)/ALPHA(J,J))
  250    CONTINUE
         FLAMDA=FLAMDA/10
 9999    RETURN
         END


C*********************************************************************

C        DESCRIPTION OF PARAMETERS
C           ARRAY  - INPUT MATRIX WHICH IS REPLACED BY ITS INVERSE
C           NORDER - ORDER OF DETERMINANT
C           DET    - DETERMINANT OF INPUT MATRIX

         SUBROUTINE MATINV(ARRAY,NORDER,DET)

         DOUBLE PRECISION ARRAY, AMAX, SAVE
         DIMENSION ARRAY(10,10)
         INTEGER IK(10),JK(10)

         DET=1.0
         DO 190 K=1,NORDER
            AMAX=0.0
  10        DO 40 I=K,NORDER
               DO 30 J=K,NORDER
                  IF (DABS(AMAX)-DABS(ARRAY(I,J))) 20,20,30
  20           AMAX=ARRAY(I,J)
               IK(K)=I
               JK(K)=J
  30           CONTINUE
  40        CONTINUE
            IF(AMAX) 60,50,60
  50        DET=0.0
            GOTO 999
  60        I=IK(K)
            IF(I-K) 10,85,70
  70        DO 80 J=1,NORDER
               SAVE=ARRAY(K,J)
               ARRAY(K,J)=ARRAY(I,J)
               ARRAY(I,J)=(-SAVE)
  80        CONTINUE
  85         J=JK(K)
            IF(J-K) 10,105,90
  90        DO 100 I=1,NORDER
               SAVE=ARRAY(I,K)
               ARRAY(I,K)=ARRAY(I,J)
               ARRAY(I,J)=(-SAVE)
  100       CONTINUE
  105       DO 120 I=1,NORDER
               IF(I-K) 120,110,120
  110          ARRAY(I,K)=-ARRAY(I,K)/AMAX
```

C-5

```
120         CONTINUE
            DO 160 I=1,NORDER
                DO 150 J=1,NORDER
                    IF (I-K) 130,160,130
130                 IF (J-K) 140,160,140
140                 ARRAY(I,J) = ARRAY(I,J) + ARRAY(I,K)*ARRAY(K,J)
150             CONTINUE
160         CONTINUE
            DO 180 J=1,NORDER
                IF (J-K) 170,180,170
170             ARRAY(K,J)=ARRAY(K,J)/AMAX
180         CONTINUE
            ARRAY(K,K)=1.0/AMAX
            DET=DET*AMAX
190     CONTINUE
        DO 240 L=1,NORDER
            K=NORDER-L+1
            J=IK(K)
            IF (J-K) 215, 215, 200
200         DO 210 I=1,NORDER
                SAVE=ARRAY(I,K)
                ARRAY(I,K)=(-ARRAY(I,J))
                ARRAY(I,J)=SAVE
210         CONTINUE
215         I=JK(K)
            IF (I-K) 240,240,220
220         DO 230 J=1,NORDER
                SAVE=ARRAY(K,J)
                ARRAY(K,J)=-ARRAY(I,J)
                ARRAY(I,J)=SAVE
230         CONTINUE
240     CONTINUE
999     RETURN
        END


C*****************************************************************

C       DESCRIPTION OF PARAMETERS
C           X       - ARRAY OF DATA POINTS FOR INDEPENDENT VARIABLE
C           I       - INDEX OF DATA POINTS
C           A       - ARRAY OF PARAMETERS

        FUNCTION FUNCTN(X,I,A,FUNC_TYPE)

        REAL A(10),X(1024)
        INTEGER I,FUNC_TYPE

        IF (FUNC_TYPE .EQ. 1) THEN
            FUNCTN=A(1)*X(I)+A(2)
        ELSEIF (FUNC_TYPE .EQ. 2) THEN
            FUNCTN=A(1)*EXP(A(2)*X(I))+A(3)
```

```
      ELSEIF (FUNC_TYPE .EQ. 3) THEN
          FUNCTN=A(1)*EXP(A(2)*X(I))+A(3)*EXP(A(4)*X(I))+A(5)
      ELSEIF (FUNC_TYPE .EQ. 4) THEN
          FUNCTN=A(1)*EXP(A(2)*X(I))+A(3)*EXP(A(4)*X(I))+
     +    A(5)*EXP(A(6)*X(I))+A(7)
       ELSEIF (FUNC_TYPE .EQ. 5) THEN
          C=1.261036E24
          FUNCTN=C*A(1)**3*(8-7*EXP(A(2)/X(I)))
      ENDIF
      RETURN
      END

C*************************************************************

C       FDERIV -- DESCRIPTION OF PARAMETERS:
C          X      - ARRAY OF DATA POINTS FOR INDEPENDENT VARIABLE
C          I      - INDEX OF DATA POINTS
C          A      - ARRAY OF PARAMETERS
C          NTERMS - NUMBER OF PARAMETERS
C          DERIV  - ARRAY OF DERIVATIVES

       SUBROUTINE FDERIV (X,I,A,DERIV,FUNC_TYPE)

       REAL A(10),DERIV(10),X(1024)
       INTEGER I,FUNC_TYPE

       IF (FUNC_TYPE .EQ. 1) THEN
          DERIV(1)=X(I)
          DERIV(2)=1.0
       ELSEIF (FUNC_TYPE .EQ. 2) THEN
          DERIV(1)=EXP(A(2)*X(I))
          DERIV(2)=X(I)*A(1)*EXP(A(2)*X(I))
          DERIV(3)=1.0
       ELSEIF (FUNC_TYPE .EQ. 3) THEN
          DERIV(1)=EXP(A(2)*X(I))
          DERIV(2)=X(I)*A(1)*EXP(A(2)*X(I))
          DERIV(3)=EXP(A(4)*X(I))
          DERIV(4)=X(I)*A(3)*EXP(A(4)*X(I))
          DERIV(5)=1.0
       ELSEIF (FUNC_TYPE .EQ. 4) THEN
          DERIV(1)=EXP(A(2)*X(I))
          DERIV(2)=X(I)*A(1)*EXP(A(2)*X(I))
          DERIV(3)=EXP(A(4)*X(I))
          DERIV(4)=X(I)*A(3)*EXP(A(4)*X(I))
          DERIV(5)=EXP(A(6)*X(I))
          DERIV(6)=X(I)*A(5)*EXP(A(6)*X(I))
          DERIV(7)=1.0
       ELSEIF (FUNC_TYPE .EQ. 5) THEN
          C=1.261036E24
          DERIV(1)=3*C*(8-7*EXP(A(2)/X(I)))*A(1)**2
          DERIV(2)=-A(1)**3*C*7/X(I)*EXP(A(2)/X(I))
```

```
        ENDIF
        RETURN
        END

C****************************************************************
C       FCHISQ -- DESCRIPTION OF PARAMETERS:
C          Y     - ARRAY OF DATA POINTS
C          SIGMAY- ARRAY OF STANDARD DEVIATIONS FOR DATA POINTS
C          NPTS  - NUMBER OF DATA POINTS
C          NFREE - NUMBER OF DEGREES OF FREEDOM
C          YFIT  - ARRAY OF CALCULATED VALUES OF Y
C          MODE  - DETERMINES METHOD OF WEIGHTING LEAST-SQARES FIT
C          1 (STATISTICAL)       WEIGHT(I) = 1.0/Y(I)
C          2 (NO WEIGHTING)      WEIGHT(I) = 1.0
C          3 (INSTRUMENTAL)      WEIGHT(I) = 1.0/SIGMAY(I)**2
C          4                     WEIGHT(I) = 1.0/DATAPOINT INDEX
C          5 LINEARLY DECREASING WEIGHT(I) = (NPTS+1-I)/NPTS

        FUNCTION FCHISQ (Y,SIGMAY,NPTS,NFREE,MODE,YFIT)

        DOUBLE PRECISION CHISQ,WEIGHT
        REAL Y(NPTS),SIGMAY(NPTS),YFIT(NPTS)

        CHISQ=0.0
        IF(NFREE) 10,10,20
10      FCHISQ=0.0
        GOTO 999
20      DO 100 I=1,NPTS
30      IF(MODE .EQ. 1) THEN
            GOTO 40
        ELSEIF(MODE .EQ. 2) THEN
            GOTO 70
        ELSEIF(MODE .EQ. 3) THEN
            GOTO 80
        ELSEIF(MODE .EQ. 4) THEN
            GOTO 85
        ELSEIF(MODE .EQ. 5) THEN
            GOTO 86
        ENDIF
40      IF(Y(I)) 60,70,50
50          WEIGHT=1.0/Y(I)
            GO TO 90
60          WEIGHT=1.0/(-Y(I))
            GO TO 90
70          WEIGHT=1.0
            GO TO 90
80          WEIGHT=1.0/(SIGMAY(I)**2)
            GOTO 90
85          WEIGHT=1.0/I
86          WEIGHT=(NPTS+1.0-I)/NPTS
```

C-8

```
90          CHISQ=CHISQ+WEIGHT*(Y(I)-YFIT(I))**2
C           PRINT*,Y(I),YFIT(I)
C           PRINT*,WEIGHT
100    CONTINUE
C           PRINT*,CHISQ,NFREE
       FCHISQ=CHISQ/NFREE
999    RETURN
       END


C*******************************************************************

C        WEIGHT_ASSIGN -- DESCRIPTION OF PARAMETERS:
C           Y     - ARRAY OF DATA POINTS
C           SIGMAY- ARRAY OF STANDARD DEVIATIONS FOR DATA POINTS
C           MODE  - DETERMINES METHOD OF WEIGHTING LEAST-SQARES FIT
C           1 (STATISTICAL)        WEIGHT(I) = 1.0/Y(I)
C             (NO WEIGHTING)       WEIGHT(I) = 1.0
C           3 (INSTRUMENTAL)       WEIGHT(I) = 1.0/SIGMAY(I)**2
C           4                      WEIGHT(I) = 1.0/DATAPOINT INDEX
C           5 LINEARLY DECREASING  WEIGHT(I) = (NPTS+1-I)/NPTS



       SUBROUTINE WEIGHT_ASSIGN(WEIGHT,Y,MODE,SIGMAY,NPTS)

       REAL SIGMAY(NPTS),Y(NPTS),WEIGHT(NPTS)
       INTEGER NPTS,I,MODE

       DO 50,I=1,NPTS
          IF(MODE .EQ. 1) THEN
             IF(Y(I) .LT. 0) THEN
                WEIGHT(I)=1.0/(-Y(I))
             ELSEIF (Y(I) .EQ. 0) THEN
                WEIGHT(I)=1.0
             ELSEIF (Y(I) .GT. 0) THEN
                WEIGHT(I)=1.0/Y(I)
             ENDIF
          ELSEIF (MODE .EQ. 2) THEN
             WEIGHT(I)=1.0
          ELSEIF (MODE .EQ. 3) THEN
             WEIGHT(I)=1.0/(SIGMAY(I)**2)
          ELSEIF (MODE .EQ. 4) THEN
             WEIGHT(I)=1.0/I
          ELSEIF (MODE .EQ. 5) THEN
             WEIGHT(I)=(NPTS+1.0-I)/NPTS
          ELSE
             WEIGHT(I)=1.0
          ENDIF
50     CONTINUE
       RETURN
       END
```

APPENDIX D - PROGRAM LASERSCAN

TABLE OF CONTENTS

PROGRAM DESCRIPTION:

General:

This program is intended for use with the Langley-Ford Model 1096 Autocorrelator for analysis of quasi-elastic light scattering data and the determination of diffusion coefficients . The program is written in several independent modules to minimize memory requirements. The program allows chaining among modules and can be interrupted by the keystroke combination <CTRL><BREAK>. The user interface is menu-driven and menus can be aborted using the <ESC> key.

Hardware requirements:

The program is intended to run on an IBM PC/AT style computer with a 80-*86 type processor and preferably equipped with a math co-processor (80-287/387). It requires 640 KB RAM and can make use of a RAM drive. A line printer is optional. The program is suitable for graphics screendumps after loading a graphics screen driver (i.e. GRAPHICS.COM). A harddisk of at least 10 MB is required. The program itself requires 2 MB of harddisk space. The program can run under all MS-DOS versions 2.* or later. A mouse is required for some of the data analysis routines.

Installation:

To install the program, place the master disk into the floppy drive and type:

B:\>install <ENTER>

The program will then extract all archived files and install itself on the harddisk in a subdirectory called "C:\LS11>". Please read the chapter on PROGRAM CONFIGURATION to determine how to set up your program.

The configured program is started by typing:

C:\LS11>LS <ENTER>

It is advisable to load a graphics screen driver before starting the program. A suitable screen driver is graphics.com for EGA adaptor boards. A commercially available program such as PIZAZZ Plus will provide greater flexibility and is recommended for usage as graphics screen handler.

If you have a floating-point math coprocessor of the type 80287 or 80387 installed in your computer, you can take advantage of the increased speed provided by the coprocessor. The data analysis programs are compiled with instruction sets that take advantage of the 80-*87 architecture to provide optimal calculation speed.

FILE FORMAT:

The file format is as follows:

Data files: <run identification.exponential sample time> (ASCII)
Data metafiles: <run identification, exponential sample time.number of components, MT> (ASCII)
Result files: <run identification, exponential sample time.res> (ASCII)

where:

1) run identification is simply the name of your run (5 characters max)
2) exponential sample time is the sample time in exponential notation, i.e., a sample time of
   $4.3 * 10^{-6}$ sec would be notated as 436.
3) number of components is the number of exponential terms fitted to the dataset (one or two)
4) MT stands for metafile.

Metafiles are automatically generated. The format for the metafiles is ASCII, with the data columns comma delimited for easy import into programs like SIGMAPLOT, QUATTRO PRO, SUPERCALC etc. For detailed information on the contents of the metafiles, please consult the respective ANALYSIS PROGRAM chapter.

Results from the analysis methods are written to files with extensions ".RES". The filename consists of the run number and the cell number: <run number, cell number,".RES">.

LASERSCAN PROGRAM

This menu allows the user to select between the main sub-programs of the ULTRASCAN program:

```
┌─────────────────────────────────────────────────────────┐
│      Laser - Autocorrelation Diffusion Analysis Program: │
├─────────────────────────────────────────────────────────┤
│                                                          │
│        ┌──────────────────────────────────────┐         │
│        │              MAIN MENU:               │         │
│        ├──────────────────────────────────────┤         │
│        │ F1: Data Acquisition                  │         │
│        │ F2: Data Analysis (PC)                │         │
│        │ F3: Data Analysis 2. Component only (PC) │       │
│        │ F4: Data Analysis (UNIX Machine)      │         │
│        │ F5: Recall old Data Analysis          │         │
│        │ F6: Program Utilities                 │         │
│        │ F7: Print Analysis                    │         │
│        │ F8: Plot Chi-Square Surfaces          │         │
│        │ F9: Exit Program                      │         │
│        └──────────────────────────────────────┘         │
│                                                          │
└─────────────────────────────────────────────────────────┘
```
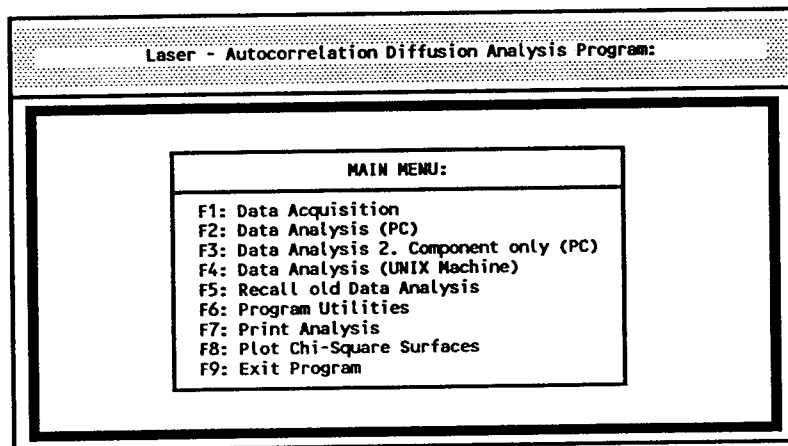
1. Data Acquisition
2. Analysis of data using the data analysis program on the PC
3. Analysis of 2. component with different starting guesses
4. Analysis of data using the data analysis program on a UNIX machine
5. Recall previously analyzed data and display results on screen
6. LaserScan Program Utilities
7. Print data analysis
8. Plot Chi-Square surfaces from desired parameters and standard deviations
9. Exit LaserScan Program

The sub-programs are selected by using the cursor keys and hitting <ENTER>. Alternatively, the indicated function keys can be pressed as a shortcut.

LASERSCAN PARAMETER SELECTION

With the LaserScan parameter selection program the basic run information can be entered and stored on harddisk. The item to be modified is selected by hitting <ENTER> on the lightbar or hitting the indicated function key as a shortcut.

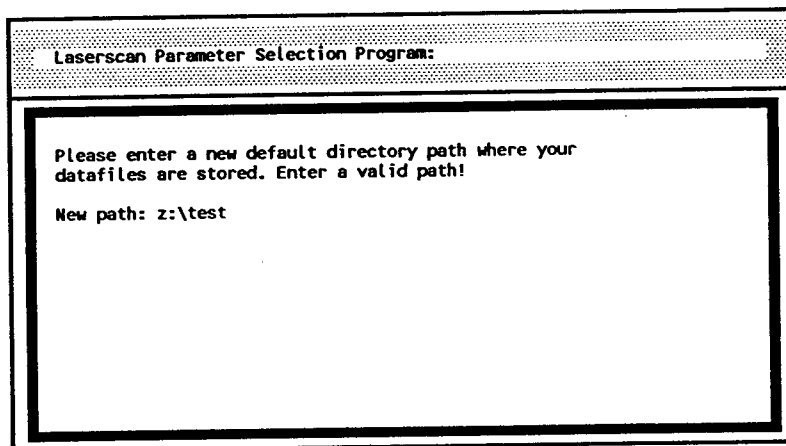

```
Laserscan Parameter Selection Program:

                    Data Parameters:

F 1: Run Identification (5 Char. max) --- TEST
F 2: Drive and Subdirectory for Data ---- D:\DATA\
F 3: Sample Time at Start of Cycle ------ 4.0 E-6
F 4: Sample Time at End of Cycle -------- 1.2 E-5
F 5: Sample Time Increment -------------- 1.0 E-6
F 6: Pause Time between Test Cycles ----- 60
F 7: Channel 1 Limit -------------------- 10000000
F 8: Run Description: None entered
F 9: Write Parameters to Disk
F10: Exit
```

1. Run identification
2. Drive and subdirectory where the run information and the data files will be stored
3. Sample time at the start of the auto-cycling data acquisition
4. Sample time at the end of the auto-cycling data acquisition
5. Increment of sample time added to each cycle in the auto-cycling data acquisition
6. Delay time between test cycles during which the computer determines the status of the data acquisition
7. Count limit of channel 1 at which data acquisition is complete
8. Complete sample description
9. Store run information on harddisk
10. Exit the parameter selection program

RUN IDENTIFICATION

Using 5 characters, the filename for the run identification can be entered. Numerical values, underscores, hyphen and alphanumeric characters can be entered. At the prompt enter the new run identification:

```
Laserscan Parameter Selection Program:


    Please enter a new run-ID. You have 5 characters to enter a unique
    identification for your run. The run-ID will be the first 5 characters
    of the filename of the file containing the ASCII data from your run.

    New Run Identification: test
```

DRIVE AND SUBDIRECTORY SELECTION

To choose a subdirectory for the location of the datafiles, select F2 and enter the new drive and subdirectory at the prompt:
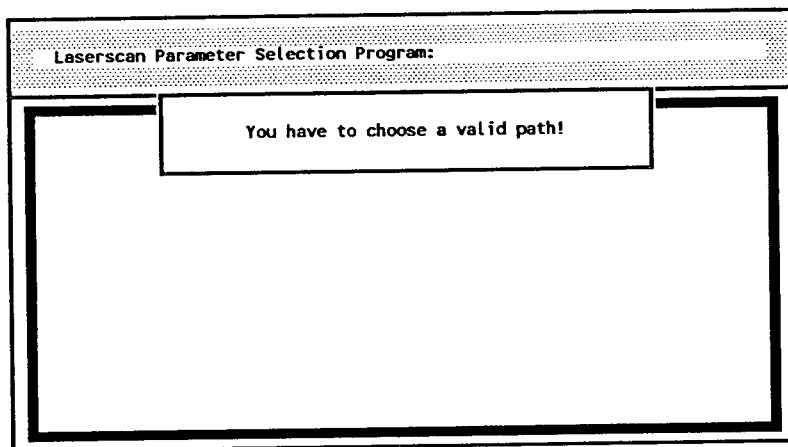
```
Laserscan Parameter Selection Program:



    Please enter a new default directory path where your
    datafiles are stored. Enter a valid path!

    New path: z:\test
```

If the subdirectory doesn't exist, an error message is printed:

```
Laserscan Parameter Selection Program:


              You have to choose a valid path!
```

SAMPLE TIME SELECTION

The program provides for a cycling mode of the data acquisition through several sample times. This feature is benefitial for capturing the entire spectrum of correlation times if the sample is not homogeneous or a single acquisition does not represent the complete correlation characteristics of the sample. The starting point for the acquisition cycle should be the shortest sample time desired for the acquisition. This time is selected by hitting F3 at the parameter selection menu. The sample time at the end of the acquisition is selected by hitting F4, and F5 will set the increment by which each successive cycle will differ from the previous cycle.

To enter a sample time, simply enter the numbers for the exponential notation of the sample time (just like on the the Langley-Ford correlator). For example, if the start sample time is $4.3 * 10^{-6}$ seconds, press 4, 3 and 6. If the increment is larger than the difference between the start and end sample time, only one acquisition will take place. Similarly, if the start sample time is the same as the end cycle time, or the end cycle time is shorter than the start cycle time, only one acquisition will be performed.

If the run information entered is new, both sample times will be set to $5.0 * 10^{-5}$ seconds, and the increment will be 0.0 seconds. If the run information is for a previously acquired run, the program will "remember" the original settings, as long as the drive and directory information is pointing to the directory where the data is stored.

```
Laserscan Parameter Selection Program:

Please enter a new sample time start (in exponential notation):

5.5 E-0
```
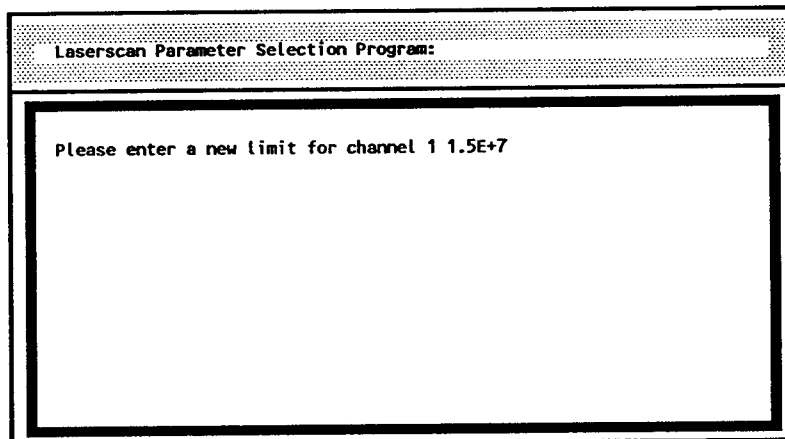
CYCLE TIME

During data acquisition, the computer communicates with the Langley-Ford Autocorrelator every so many seconds to check on the progress of data acquisition. The progress is evaluated as the number of counts for the first channel of the correlation spectrum. If a limiting value has been reached, the program will write the current acquisition state to a datafile and cycle to the next sample time.

This cycle time is set according to the number of counts correlated by the correlator. If there are a lot of counts, the cycle time may be set to a shorter time, however, if acquisition takes longer to reach a desired value for channel one, the cycle time should be set to a longer time, because each communication with the correlator takes approximately 7 seconds time away from the actual acquisition. A good default value for the cycle time is 60 seconds. To enter the new cycle time delay, select F6 from the parameter selection program and enter the desired seconds at the prompt.

```
 Laserscan Parameter Selection Program:


    Please enter a new pause time for each check cycle: 90
```

CHANNEL 1 LIMIT
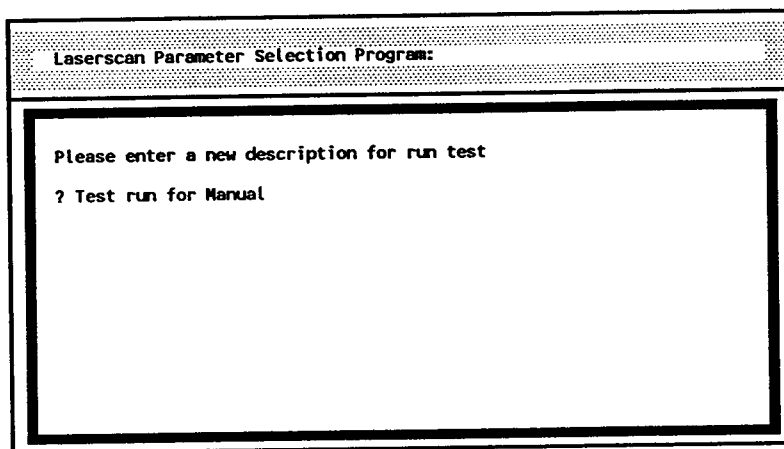
The limit for channel 1 determines the status of acquisition, at which the data is written to a file and the auto-cycling mode will reset the correlator to a new sample time. A value at which the data is usually long enough averaged out is $1.0 * 10^7$ counts total. Higher counts usually don't add accuracy to the data. Lower counts can be used when sample counts are very low because of low sample concentration or strong overflow propensity, requiring that the laser intensity is diminuished. To enter a new count for channel one, select F7 and enter the new channel 1 limit at the prompt:

```
Laserscan Parameter Selection Program:


  Please enter a new limit for channel 1 1.5E+7
```
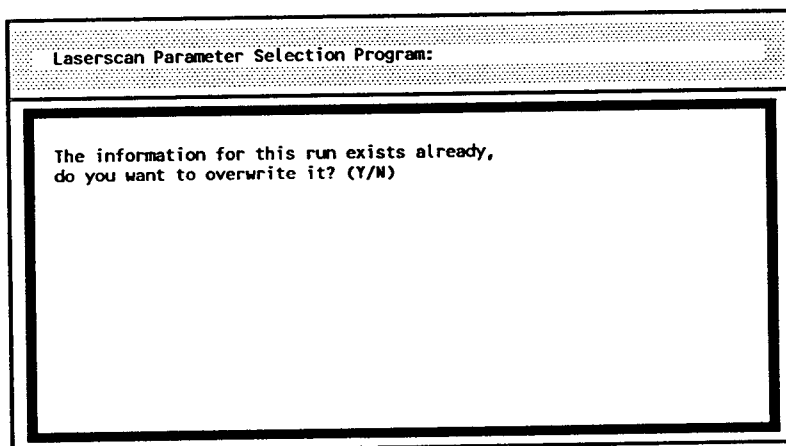
SAMPLE DESCRIPTION

Since five characters available for the file name are usually not explicit enough for a complete sample description, a more complete sample description can be entered by selecting F8 from the parameter selection menu:

```
Laserscan Parameter Selection Program:


    Please enter a new description for run test

    ? Test run for Manual




```
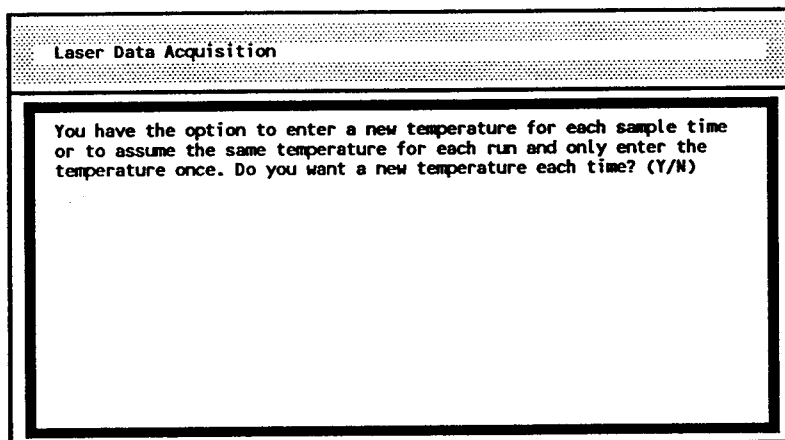
After all parameters for the new run have been entered, the setup should be saved on the harddrive by selecting F9 from the parameter selection program. If a particular configuration for a run of the same file name has been saved previously, the program responds with an error message and warns the user that the origianl data will be overwritten. At this point answering "N" will abort the overwriting and allow the user to enter a different filename for the data acquisition.

```
Laserscan Parameter Selection Program:

    The information for this run exists already,
    do you want to overwrite it? (Y/N)




```

DATA ACQUISITION

Data acquisition normally cycles through several sample times to capture the entire correlation spectrum of the sample. Since temperature control over longer periods of time might be unstable (due to equipment limitations - what else is new?), the program allows to enter a new temperature value used for the viscosity correction before each new sample time cycle begins. However, if your system is trouble free, you might want to skip this option and walk away from your experiment while data acquisition takes place (have a cup of coffe etc.).

```
┌─────────────────────────────────────────────────────────┐
│   Laser Data Acquisition                                 │
│ ┌─────────────────────────────────────────────────────┐ │
│ │ You have the option to enter a new temperature for each sample time │ │
│ │ or to assume the same temperature for each run and only enter the   │ │
│ │ temperature once. Do you want a new temperature each time? (Y/N)     │ │
│ │                                                     │ │
│ │                                                     │ │
│ │                                                     │ │
│ │                                                     │ │
│ │                                                     │ │
│ └─────────────────────────────────────────────────────┘ │
└─────────────────────────────────────────────────────────┘
```
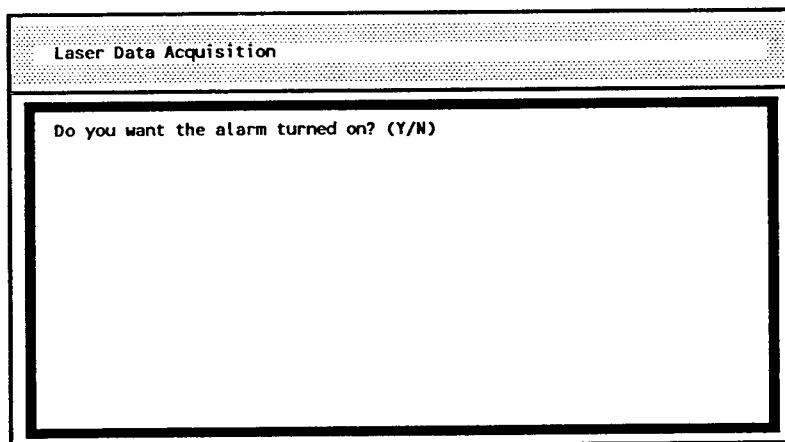
If you select to enter a new temperature for each sample time, you can turn on an alarm, which signals to you the end of each acquisition cycle and reminds you to enter a new temperature before the new sample time will be set.
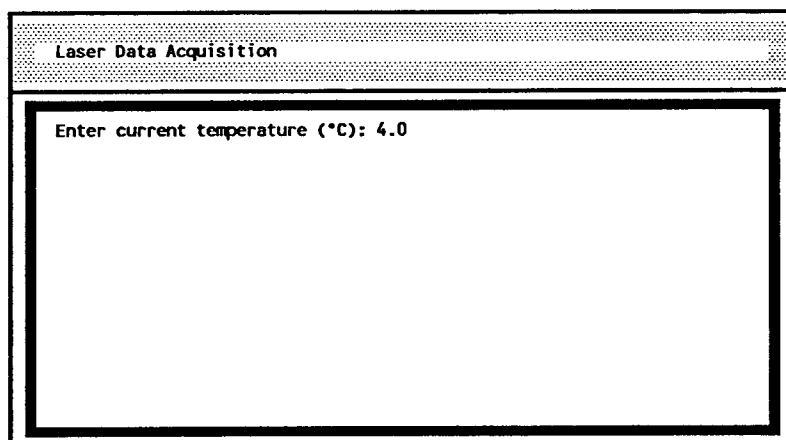
```
┌─────────────────────────────────────────────────────────┐
│   Laser Data Acquisition                                 │
│ ┌─────────────────────────────────────────────────────┐ │
│ │ Do you want the alarm turned on? (Y/N)              │ │
│ │                                                     │ │
│ │                                                     │ │
│ │                                                     │ │
│ │                                                     │ │
│ │                                                     │ │
│ │                                                     │ │
│ └─────────────────────────────────────────────────────┘ │
└─────────────────────────────────────────────────────────┘
```

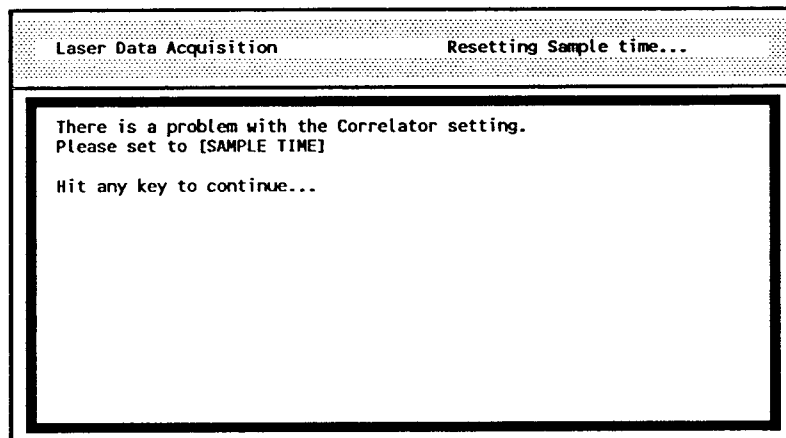## TEMPERATURE CORRECTION

The temperature entered during data acquisition will be used to calculate a viscosity correction for the diffusion coefficient calculation. The correction is based on water at 20°C and is taken from the empirical relationship published in the 62nd edition of the CRC Handbook for Chemistry and Physics (pp. F42).

```
┌─────────────────────────────────────────────────────────┐
│    Laser Data Acquisition                                │
├─────────────────────────────────────────────────────────┤
│                                                          │
│   Enter current temperature (°C): 4.0                    │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
└─────────────────────────────────────────────────────────┘
```

Data Acquisition is started by establishing a serial RS232 communication link between the computer and Langley-Ford autocorrelator. The sample time will be reset to the start value entered under F3 of the parameter selection menu. If communication does not succeed, an error message will be printed:

```
┌─────────────────────────────────────────────────────────┐
│    Laser Data Acquisition          Resetting Sample time...│
├─────────────────────────────────────────────────────────┤
│                                                          │
│   There is a problem with the Correlator setting.        │
│   Please set to [SAMPLE TIME]                            │
│                                                          │
│   Hit any key to continue...                             │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
│                                                          │
└─────────────────────────────────────────────────────────┘
```

If this situation occurs, please check following items:

1. Make sure that the correlator is connected via a null modem serial connection to the computer.
2. Verify that the serial port defined in the Program Configuration is indeed the one connected to the serial interface.
3. Make sure the screen of the Langley-Ford correlator is set to "SAMPLE TIME".

CYCLING MODE

Once communication between the computer and the correlator has been established, the computer will reset the system clock of the correlator to "0", reset the sample time to the correct time and start acquisition. During acquisition, the computer will check the status of acquisition on channel 1 of the correlator by requesting channel 1 counts in intervals defined by the pause time between each test cycle with F6 in the parameter selection menu. The status of data acquisition will continuously be displayed and scroll through the window as more test cycles have been completed. The sample time and the total elapsed time will also be displayed.

```
  Laser Data Acquisition, S-Time: 4.0 E-6 SEC, Elapsed Time: 535 SEC

      Cycle: 28,   Channel-1: 5.024000E+5,   Elapsed Time: 354 seconds
      Cycle: 29,   Channel-1: 5.031360E+5,   Elapsed Time: 367 seconds
      Cycle: 30,   Channel-1: 5.040320E+5,   Elapsed Time: 379 seconds
      Cycle: 31,   Channel-1: 5.044800E+5,   Elapsed Time: 392 seconds
      Cycle: 32,   Channel-1: 5.086400E+5,   Elapsed Time: 405 seconds
      Cycle: 33,   Channel-1: 5.137600E+5,   Elapsed Time: 417 seconds
      Cycle: 34,   Channel-1: 5.167040E+5,   Elapsed Time: 430 seconds
      Cycle: 35,   Channel-1: 5.257600E+5,   Elapsed Time: 443 seconds
      Cycle: 36,   Channel-1: 5.397120E+5,   Elapsed Time: 455 seconds
      Cycle: 37,   Channel-1: 5.524800E+5,   Elapsed Time: 468 seconds
      Cycle: 38,   Channel-1: 5.644480E+5,   Elapsed Time: 480 seconds
      Cycle: 39,   Channel-1: 5.743680E+5,   Elapsed Time: 493 seconds
      Cycle: 40,   Channel-1: 5.801280E+5,   Elapsed Time: 506 seconds
      Cycle: 41,   Channel-1: 5.851840E+5,   Elapsed Time: 519 seconds
      Cycle: 42,   Channel-1: 5.938880E+5,   Elapsed Time: 531 seconds
```

D-17

DATA ANALYSIS (PC)

Once data acquisition is completed, the data can be analyzed in two different ways. The calculations can be performed on the PC or the data can be transferred to a UNIX based mainframe for faster data processing. In either case, the data is analyzed in the order as the cycling mode has been set up. To analyze the data, various parameters for the analysis have to be set:

```
Selection of Laser Parameters:

Current Selection for Run ID 20_0:

        Laser Parameters:

    F1: Refractive Index --- 1.33
    F2: Scattering Angle --- 90
    F3: Laser Wavelength --- 488
    F4: Relative Viscosity - 1
    F5: Density ----------- 1
    F6: V-bar ------------- .6
    F7: Fitting Tolerance -- .00001
    F8: Weighing Method ---- 1.0
    F9: Exit
```

1. Refractive Index (the refractive index of the solvent)
2. Scattering Angle (the angle between the laser beam and the photomultiplier tube)
3. Laser Wavelength
4. Relative Viscosity (the viscosity of the sample at 20°C relative to water at 20°C)
5. Density (not required for diffusion coefficient calculation)
6. V-bar (not required for diffusion coefficient calculation)
7. Fitting Tolerance (the magnitude of the statistical F-test at which fitting of the data is considered to be converged)
8. Weighing Method (the mode by which the data is weighted, several different schemes are available for the fitting of the data)
9. Exit (the analysis parameter selection menu)

REFRACTIVE INDEX

At the prompt, enter the refractive index of your solvent:

```
┌─────────────────────────────────────────────────────────────┐
│  Selection of Laser Parameters:                               │
│                                                               │
│  ┌──────────────────────────────────────────────────────┐    │
│  │                                                        │    │
│  │  Current Selection for Run ID 20_0:                    │    │
│  │                                                        │    │
│  │                                                        │    │
│  │  Please enter the refractive index: 1.33               │    │
│  │                                                        │    │
│  └──────────────────────────────────────────────────────┘    │
└─────────────────────────────────────────────────────────────┘
```

At the prompt, enter the value for the viscosity of the solvent at 20°C relative to the viscosity of water at 20°C:

```
┌─────────────────────────────────────────────────────────────┐
│  Selection of Laser Parameters:                               │
│                                                               │
│  ┌──────────────────────────────────────────────────────┐    │
│  │                                                        │    │
│  │  Current Selection for Run ID 20_0:                    │    │
│  │                                                        │    │
│  │                                                        │    │
│  │  Please enter the viscosity relative to water: 1.001   │    │
│  │                                                        │    │
│  └──────────────────────────────────────────────────────┘    │
└─────────────────────────────────────────────────────────────┘
```

FITTING TOLERANCE

The fitting tolerance determines the point at which the fitting routine (performing a nonlinear fit) will consider the fitting to be converged. This is the value the statistical F-test has to attain, before the iterative process of the fitting routine is abandoned. This value should be less than $1.0*10^{-4}$:

```
Selection of Laser Parameters:

  Current Selection for Run ID 20_0:


  Please enter the fitting tolerance: 0.00001
```

WEIGHING METHOD

The weighing method determines the importance (weight) given to each datapoint. Five different methods are available:

1. 1.0/datapoint
2. no weighting - weight = 1.0
3. 1.0/(standard deviation of datapoint)^2
4. 1.0/datapoint index
5. linearly decreasing weighting

The best method to use for data acquired with the Langley-Ford autocorrelator is weighing method 2, no weighting.

```
  Selection of Laser Parameters:


    Current Selection for Run ID 20_0:


    Please enter the weighing method (1,2,3,4,5): 2

    Available options are:

    (1): weight = 1.0/datapoint
    (2): weight = 1.0
    (3): weight = 1.0/sigma**2(datapoint)
    (4): weight = 1.0/datapoint index
    (5): weight = linearly decreasing
```

DATA ANALYSIS

The method for data analysis is a non-linear, least squares fitting routine, which minimizes the Chi-square by steepest gradient descent on the Chi-square surface with respect to all parameters. The function to be minimized is:

$$F(\tau) = \sum_{i=1}^{n} [a_i * \exp(b_i * \tau)] + c$$

with parameters a_i, b_i and c to be fitted. The parameters a_i designate the amplitude, or relative amount of the ith component and parameter b_i designates the correlation frequency of the ith component. Parameters b_i are proportional to the diffusion coefficient. The baseline is denoted by parameter c.

The number of components is designated by n, where n can be one or two. The routine employed is the Levenberg-Marquardt fitting routine (Bevington), modified for use with the scattering data. Initial guesses for all paramters are required. The baseline is estimated by averaging the last 16 delayed channels from the autocorrelator. This value is subtracted from all datapoints and then fitted by a linear least squares fitting routine to give the initial guesses for a single component fit. These parameters are imported to the routine non-linear least squares routine to obtain optimized values for a_1, b_1 and c.
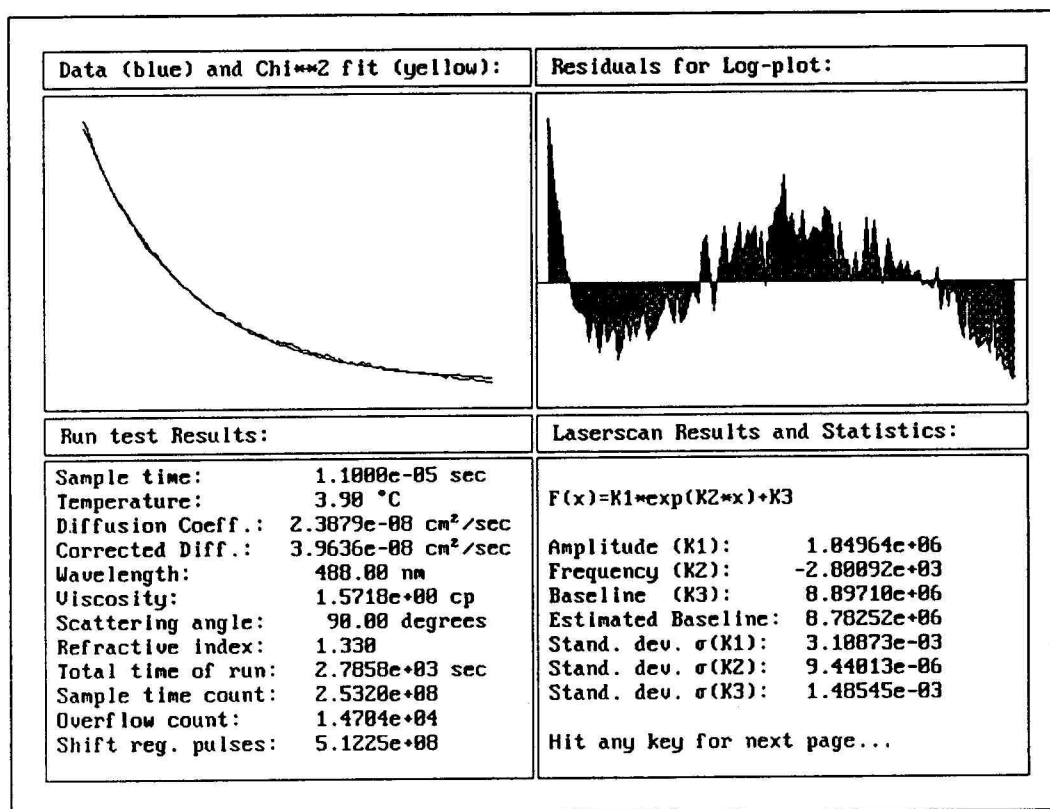
During fitting, the status of the fitting process is displayed on the screen:

```
Please wait, calculating...

Iteration:   60, F-Test:    .1335410E+01, X-Square:   .8115870E+09
```
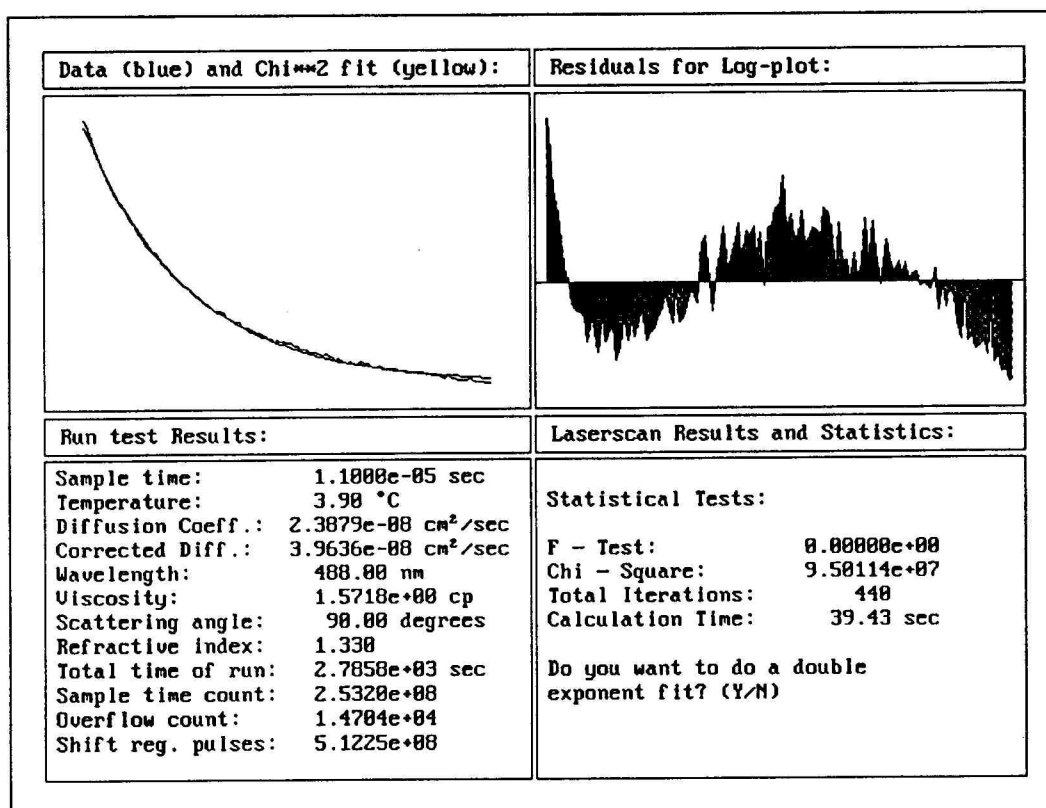
SINGLE COMPONENT FIT:

Once the single component fit is completed, the fitted data and calculated results will be displayed on the screen:

Data (blue) and Chi**2 fit (yellow):	Residuals for Log-plot:

Run test Results:	Laserscan Results and Statistics:

Sample time:	1.1000e-05 sec	F(x)=K1*exp(K2*x)+K3	
Temperature:	3.90 °C		
Diffusion Coeff.:	2.3879e-08 cm²/sec	Amplitude (K1):	1.04964e+06
Corrected Diff.:	3.9636e-08 cm²/sec	Frequency (K2):	-2.80092e+03
Wavelength:	488.00 nm	Baseline (K3):	8.89710e+06
Viscosity:	1.5718e+00 cp	Estimated Baseline:	8.78252e+06
Scattering angle:	90.00 degrees	Stand. dev. σ(K1):	3.10873e-03
Refractive index:	1.330	Stand. dev. σ(K2):	9.44013e-06
Total time of run:	2.7858e+03 sec	Stand. dev. σ(K3):	1.48545e-03
Sample time count:	2.5320e+08		
Overflow count:	1.4704e+04	Hit any key for next page...	
Shift reg. pulses:	5.1225e+08		

In the left/top panel the data and the overlayed fitted, analytical function is displayed. In the right/top panel the distribution of the Chi-square residuals about the fit is displayed in a magnified format. In the left/bottom panel, correlator specific data for the run are displayed, as well as parameters used for the calculation of the diffusion coefficient. The viscosity is indicated as a temperature corrected viscosity and the diffusion coefficient is calculated both temperature corrected and uncorrected. In the right/bottom panel the fitting results are displayed, showing both the estimated and calculated baseline, the amplitude, frequency and standard deviations for each calculated parameter.
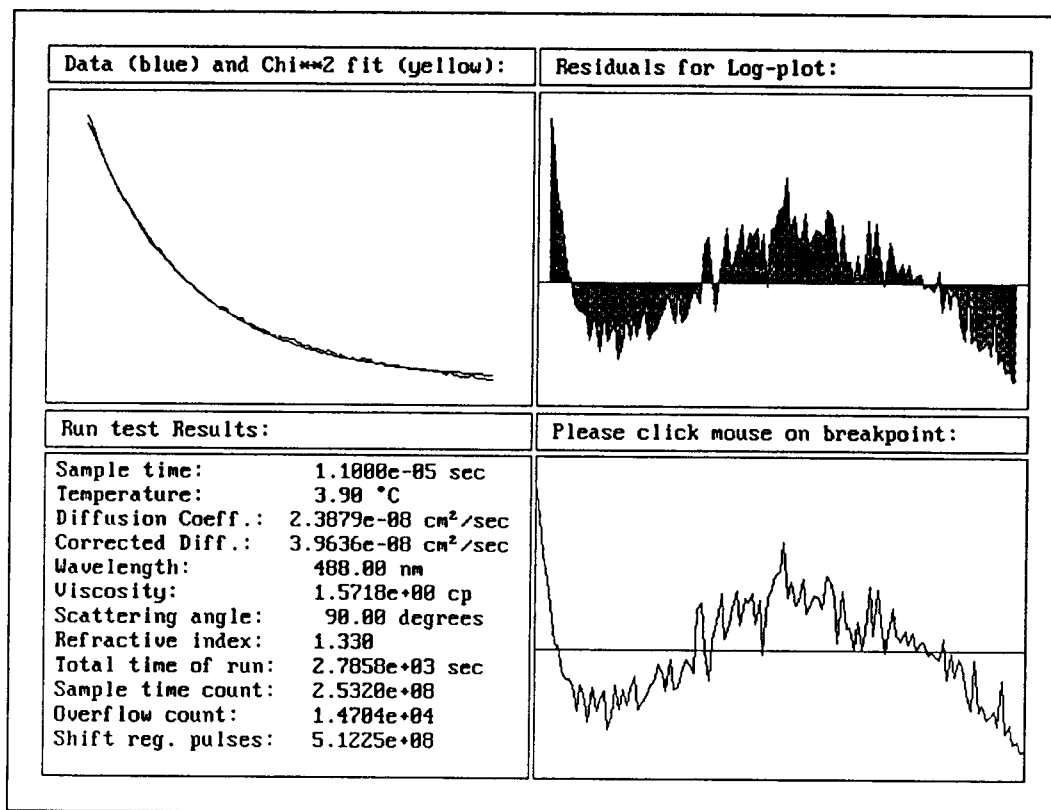
On the next page, the statistical results from the analysis are displayed in the left/bottom panel:

Data (blue) and Chi**2 fit (yellow):	Residuals for Log-plot:

Run test Results:	Laserscan Results and Statistics:

Run test Results:		Laserscan Results and Statistics:	
Sample time:	1.1000e-05 sec	**Statistical Tests:**	
Temperature:	3.90 °C		
Diffusion Coeff.:	2.3879e-08 cm²/sec	F – Test:	0.00000e+00
Corrected Diff.:	3.9636e-08 cm²/sec	Chi – Square:	9.50114e+07
Wavelength:	488.00 nm	Total Iterations:	440
Viscosity:	1.5718e+00 cp	Calculation Time:	39.43 sec
Scattering angle:	90.00 degrees		
Refractive Index:	1.330	Do you want to do a double	
Total time of run:	2.7858e+03 sec	exponent fit? (Y/N)	
Sample time count:	2.5320e+08		
Overflow count:	1.4704e+04		
Shift reg. pulses:	5.1225e+08		

Consistent "runs" of residuals in the right/top panel suggest a poor fit with the particular model. In a case where "runs" are present, a fit with more degrees of freedom, i.e., a multicomponent fit will fare better.

DOUBLE COMPONENT INITIAL GUESSES

To select a double component fit, anser "Y" and select a break point with the mouse for the calculation of the initial parameters of the second component fit. The initial guesses for the parameters are found by dividing the whole spectrum into two parts and fitting each part individually to a linear least squares fit, after the optimized baseline has been subtracted. The idea is that slower components will contribute mainly to the longer correlation spectrum (on the right side), while faster diffusing molecules will contribute primarily to the left side of the spectrum. If "runs" are obvious from the right/top panel, the breakpoint is best selected in between the two apparent components.

Data (blue) and Chi2 fit (yellow):**		**Residuals for Log-plot:**

Run test Results:

Sample time:	1.1000e-05 sec
Temperature:	3.90 °C
Diffusion Coeff.:	2.3879e-08 cm²/sec
Corrected Diff.:	3.9636e-08 cm²/sec
Wavelength:	488.00 nm
Viscosity:	1.5718e+00 cp
Scattering angle:	90.00 degrees
Refractive index:	1.330
Total time of run:	2.7858e+03 sec
Sample time count:	2.5320e+08
Overflow count:	1.4704e+04
Shift reg. pulses:	5.1225e+08

Please click mouse on breakpoint:

After the breakpoint has been selected with the mouse, initial guesses for the five parameter - double component fit are calculated and analyzed by the Levenberg-Marquardt fitting routine.

DOUBLE COMPONENT FIT

The double component fit is displayed similarly to the single component fit:

Data (blue) and Chi**2 fit (yellow):	Residuals for Log-plot:

Run test Results:	Laserscan Results and Statistics:
Sample time: 1.1000e-05 sec Temperature: 3.90 °C Diffusion Coeff. for 2 components: 1. Component: 7.0429e-08 cm²/sec Corrected(20°,W): 1.1690e-07 cm²/sec Relative Amount: 60.12 % 2. Component: 1.3911e-08 cm²/sec Corrected(20°,W): 2.3090e-08 cm²/sec Relative Amount: 39.88 %	F(x)=K1*exp(K2*x)+K3*exp(K4*x)+K5 Amplitude-1 (K1): 7.47960e+05 Frequency-1 (K2): -4.13052e+03 Amplitude-2 (K3): 4.96158e+05 Frequency-2 (K4): -8.15830e+02 Baseline (K5): 8.73488e+06 Stand. dev. σ(K1): 1.73867e-03 Stand. dev. σ(K2): 1.22334e-05 Stand. dev. σ(K3): 5.72165e-04 Stand. dev. σ(K4): 1.55595e-06 Stand. dev. σ(K5): 8.15044e-04 Hit any key for next page...

Generally, the distribution of the residuals about the fitted line becomes more random as more components are fitted and the "runs" from poorly fitted data disappears.

Again, the statistical results for the double component fit are displayed on the right/bottom panel of the next page:

Data (blue) and Chi**2 fit (yellow):	Residuals for Log-plot:

Run test Results:	Laserscan Results and Statistics:

Sample time: 1.1000e-05 sec
Temperature: 3.90 °C

Diffusion Coeff. for 2 components:

1. Component: 7.0429e-08 cm²/sec
Corrected(20°,W):   1.1690e-07 cm²/sec
Relative Amount: 60.12 %
2. Component: 1.3911e-08 cm²/sec
Corrected(20°,W):   2.3090e-08 cm²/sec
Relative Amount: 39.88 %

Statistical Tests:

F - Test: 0.00000e+00
Chi - Square: 1.61246e+07

Total Iterations: 1412
Calculation Time: 246.40 sec

Hit any key to continue...

If the fit is not satisfactory, it is possible to repeat the 2 component fit by selecting (F3) from the main menu. In order to use this feature, a single component fit has to exist already. The 2 component fitting routine can be run as many times as necessary, each time with a different initial guess for the parameters.

RECALL OLD ANALYSIS

Once a dataset has been analyzed, the data can be recalled without having to re-calculate the parameters for the fit. To recall a previously analyzed dataset, select RECALL OLD DATA ANALYSIS from the MAIN MENU (F3) and enter the LaserScan parameters for the desired data. The program will read in the results file and display the fitted data. If the fitted data is unavailable, following error message will be printed:

```
Following Error occurred during Execution:


   The analyzed data for D:\DATA\test.125 (double component)
   is not available.

   Hit any key to continue or <CTRL> <BREAK> to abort...
```
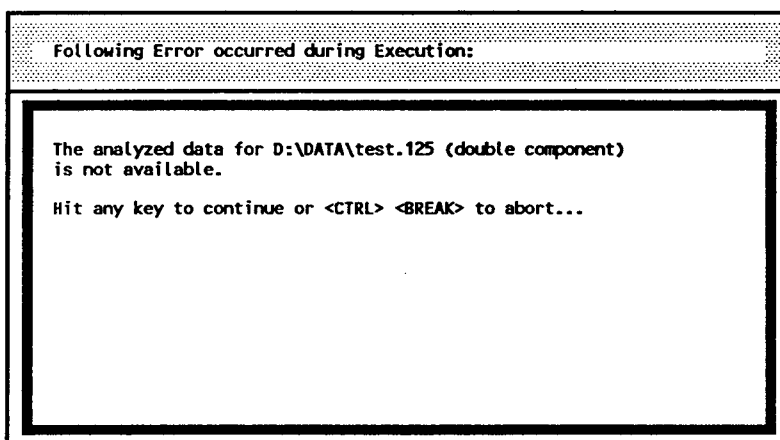
MAINFRAME DATA ANALYSIS

The LaserScan program provides for a method to analyze the data on a mainframe. The data is transferred to the mainframe over a serial communication link using KERMIT. A script is invoked which will logon to a UNIX machine and transfer the datafiles for all included runs to the mainframe. On the mainframe, a Levenberg-Marquardt fitting routine is invoked and used to analyze the data. If a mainframe is available, data processing is usually much quicker than on a PC and will save time. To write the datafiles for the mainframe, select DATA ANALYSIS - WRITE FILE FOR UNIX MACHINE (F4):

```
Please wait, loading data...
Writing single component file for run test.115 to: out1.115
Writing double component file for run test.115 to: out2.115
Please wait, loading data...
Writing single component file for run test.125 to: out1.125
Writing double component file for run test.125 to: out2.125
```

After the files are written, select KERMIT FILE TRANSFER to transfer the files to the mainframe. A warning message will be displayed, reminding the user to connect the computer to the modem or serial connection:

```
Please make sure that the UNIX machine is connected
to the proper serial communication line (COM port)!

Hit any key to invoke the KERMIT script...
```

The next step is the actual transfer of data:

```
MS-DOS Kermit: 3.02 dev 19 Sept 1990


            File name: OUT1.115
  KBytes transferred: 2
 Percent transferred: 46%
             Sending: In progress

   Number of packets: 17
       Packet length: 88
   Number of retries: 2
          Last error:
        Last message: Remote name is out1.115



    X: cancel file, Z: cancel group, E: exit nicely, C: exit abruptly, Enter: retry
```

After the data has been transferred, the Levenberg-Marquardt routine will calculate the fit for each individual datafile in the order they were acquired:

```
Calculating fit for: out1.115
Total elapsed time:    6.390000
Calculation time for last fit:    6.400000
Total number of iterations:        391

Calculating fit for: out2.115
Total elapsed time:    51.73000
Calculation time for last fit:   45.33000
Total number of iterations:       1360

-------------------------------------------------

Calculating fit for: out1.125
Total elapsed time:    57.27000
Calculation time for last fit:    5.550000
Total number of iterations:        340

Calculating fit for: out2.125
```

D-29

The last step is the transfer of the results back to the PC. Once the analyzed data is successfully transferred to the PC, both single and double component results will be displayed. If the second component analysis yield unsatisfactory results, the second component analysis can be repeated by selecting (F3 - DATA ANALYSIS 2. COMPONENT ONLY (PC)) from the MAIN MENU.

PROGRAM UTILITITES

The utility program from the LaserScan program allows the user to select certain parameters necessary for proper program functioning and to chain to shell to additional programs:

```
┌────────────────────────────────────────────────────────────┐
│  Laserscan Data Analysis Program:                          │
├────────────────────────────────────────────────────────────┤
│  ┌──────────────────────────────┐                          │
│  │       UTILITY MENU:          │                          │
│  ├──────────────────────────────┤                          │
│  │  F1: Desktop Utility         │                          │
│  │  F2: Program Configuration   │                          │
│  │  F3: Escape to DOS Shell     │                          │
│  │  F4: Exit Utility Program    │                          │
│  └──────────────────────────────┘                          │
│                                                            │
│                                                            │
└────────────────────────────────────────────────────────────┘
```
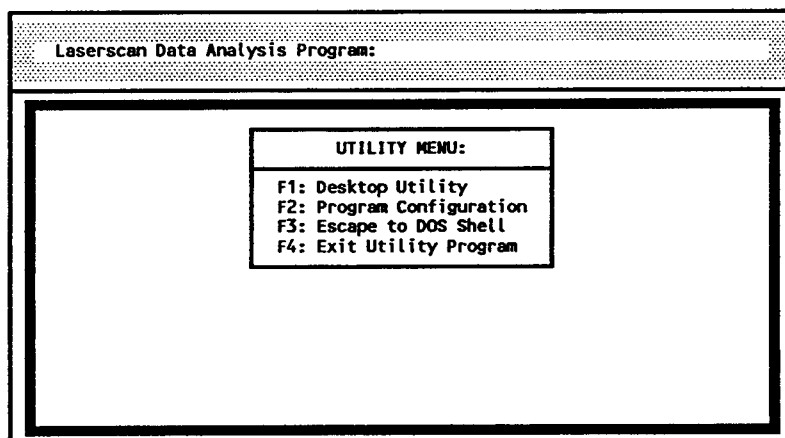
1. Shell to a desktop utility to organize and copy files on hard- and floppy disks
2. Setup of hardware and software to customize the environment
3. Escape to a disk operating command shell
4. Exit the utility program

PROGRAM CONFIGURATION

The program configuration menu lets the user control the hardware setups for the serial communication between the computer and the correlator. Further, a graphics mode suitable for the monitor attached to the PC can be selected. The best results for graphics are obtained with a VGA style monitor. Also, the datapath to the desktop utility can be defined in the configuration program:

```
┌──────────────────────────────────────────────────────┐
│  Laserscan Configuration Program:                      │
├──────────────────────────────────────────────────────┤
│  ┌──────────────────────────────────────────┐          │
│  │        Configuration Parameters:          │          │
│  │  F1: Serial Port ------ 2                 │          │
│  │  F2: Baud Rate -------- 9600              │          │
│  │  F3: Databits --------- D7                │          │
│  │  F4: Stopbits --------- S1                │          │
│  │  F5: Parity ----------- PE                │          │
│  │  F6: Graphics Monitor - VGA               │          │
│  │  F7: Desktop Utility -- D:\UTIL\PCSHELL.EXE│          │
│  │  F8: Exit Configuration Program           │          │
│  └──────────────────────────────────────────┘          │
└──────────────────────────────────────────────────────┘
```

COM PORT SELECTION

The COM port setting defines which serial port is used for data communication with the correlator. This setting depends on the configuraton and availability of the hardware on the PC. Consult your computers USER MANUAL to decide which COM port you should use.

```
┌──────────────────────────────────────────────────────┐
│  Laserscan Configuration Program:                      │
├──────────────────────────────────────────────────────┤
│  ┌──────────────────────┐                              │
│  │   COM Port setting:   │                              │
│  │  F1:  COM 1           │                              │
│  │  F2:  COM 2           │                              │
│  │  F3:  COM 3           │                              │
│  │  F4:  COM 4           │                              │
│  │  F5:  COM 5           │                              │
│  │  F6:  COM 6           │                              │
│  │  F7:  COM 7           │                              │
│  │  F8:  COM 8           │                              │
│  │                       │                              │
│  │  DEFAULT = COM 1      │                              │
│  └──────────────────────┘                              │
└──────────────────────────────────────────────────────┘
```

BAUD RATE SELECTION

The baud rate determines the speed of communication between the computer and the correlator. Like all communication settings, this setting has to agree with the communication paramters selected on the correlator. If communication lines are short in distance and in an environment which is electronically not noisy, the highest data transfer rate available should be selected. The Langley-Ford correlator supports baud rates of up to 9600 baud.

```
Laserscan Configuration Program:

Baud Rate:

F1:   300
F2:   600
F3:  1200
F4:  2400
F5:  4800
F6:  9600
F7: 19600

DEFAULT = 9600
```

DATABIT SELECTION

Databits are the number of bits sefining the length of one word transmitted. In an 8-bit machine, the transfer is usually supported with 7 databits and 1 stopbit. Again, the databit setting has to agree with the setting on the correlator:

```
Laserscan Configuration Program:

Databits

F1:  5
F2:  6
F3:  7
F4:  8

DEFAULT = 7
```

STOPBIT SELECTION

If you have selected 7 databits, you should use 1 stopbit, similarly, if you have selected 6 databits, you should use 2 stopbits. Again, this setting has to agree with the setting selected on the correlator:

```
Laserscan Configuration Program:

   Stopbits:

   F1:    1
   F2:    2

   DEFAULT = 1
```
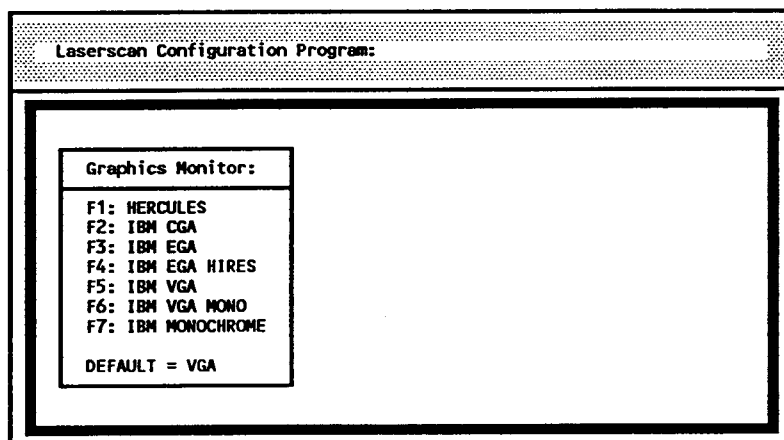
PARITY SELECTION

Select parity between even (default), odd, space, mark and none:

```
Laserscan Configuration Program:

    Parity:

   F1:  Even
   F2:   Odd
   F3: Space
   F4:  Mark
   F5:  None

   DEFAULT = Even
```

GRAPHICS MONITOR SELECTION

Select the graphics monitor connected to your system. It is preferrable to have a high resolution graphics monitor like VGA or VGA MONO to get best resolution on the graphics applications:
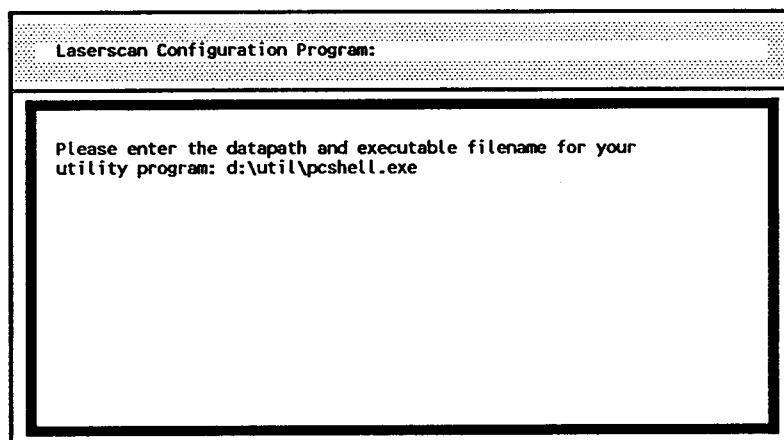
```
 Laserscan Configuration Program:

    ┌─────────────────────┐
    │ Graphics Monitor:   │
    │                     │
    │ F1: HERCULES        │
    │ F2: IBM CGA         │
    │ F3: IBM EGA         │
    │ F4: IBM EGA HIRES   │
    │ F5: IBM VGA         │
    │ F6: IBM VGA MONO     │
    │ F7: IBM MONOCHROME  │
    │                     │
    │ DEFAULT = VGA       │
    └─────────────────────┘
```

DESKTOP UTILITY PATH

Here you enter the directory path where your desktop utility or program editor resides. Any executable program can be entered and used for chaining in conjunction with LaserScan.

```
 Laserscan Configuration Program:

    Please enter the datapath and executable filename for your
    utility program: d:\util\pcshell.exe
```

D-35

DOS SHELL

Use the DOS SHELL command to exit LaserScan temporarily, enter EXIT at the DOS prompt to return to LaserScan:

```
LaserScan Program Shell, Version 1.10
Please enter 'EXIT' to return to LaserScan Program.
```

```
Microsoft(R) MS-DOS(R) Version 5.00
          (C)Copyright Microsoft Corp 1981-1991.

C:\LASER>
```

LASER PRINTING PROGRAM

This routine allows the user to print out data results and to obtain graphics screen dumps of the graphics displayed in the analysis. These graphics are displayed in a full-page mode to take advantage of the entire screen for optimal resolution. A suitable graphics driver is required for capturing the display memory and printing it to a printer or file. This graphics driver should be loaded before LaserScan is started.

```
Laser - Printing Program



      MAIN MENU:

   F1: Print Graphics
   F2: Print Datafiles
   F3: Exit Program
```
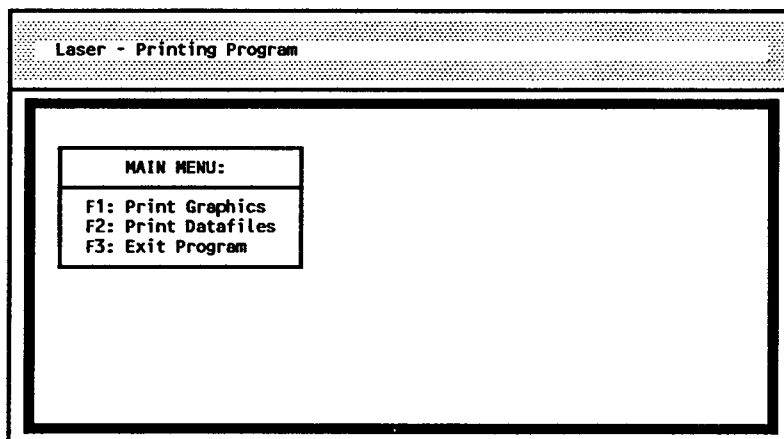
When printing the graphics, first the single component fit and data are shown in an overlayed fashion, then the residuals are shown and finally the same information for the double component fit are displayed.

The program will look for a screen driver activated by the <SHIFT> <PRINT SCREEN> sequence.

SINGLE COMPONENT FIT
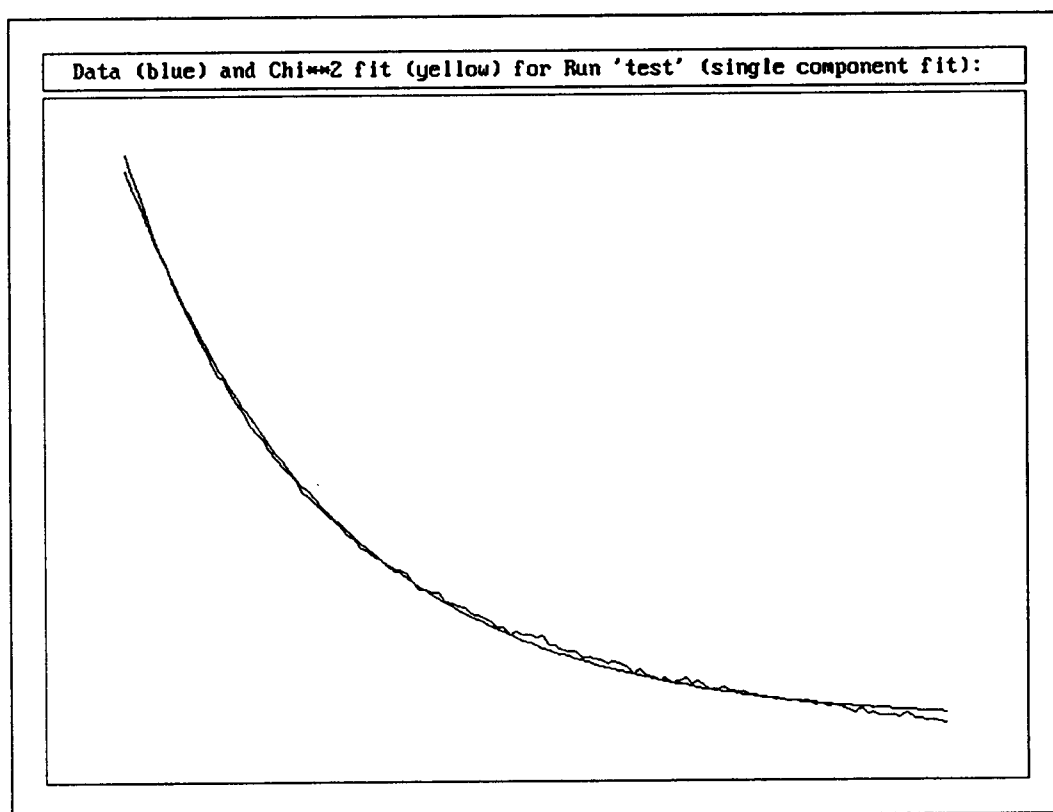
Data (blue) and Chi**2 fit (yellow) for Run 'test' (single component fit):

SINGLE COMPONENT RESIDUALS

Residuals for Log-plot for Run 'test' (single component fit):

DOUBLE COMPONENT FIT

Data (blue) and Chi**2 fit (yellow) for Run 'test' (double component fit):

DOUBLE COMPONENT RESIDUALS

Residuals for Log-plot for Run 'test' (double component fit):
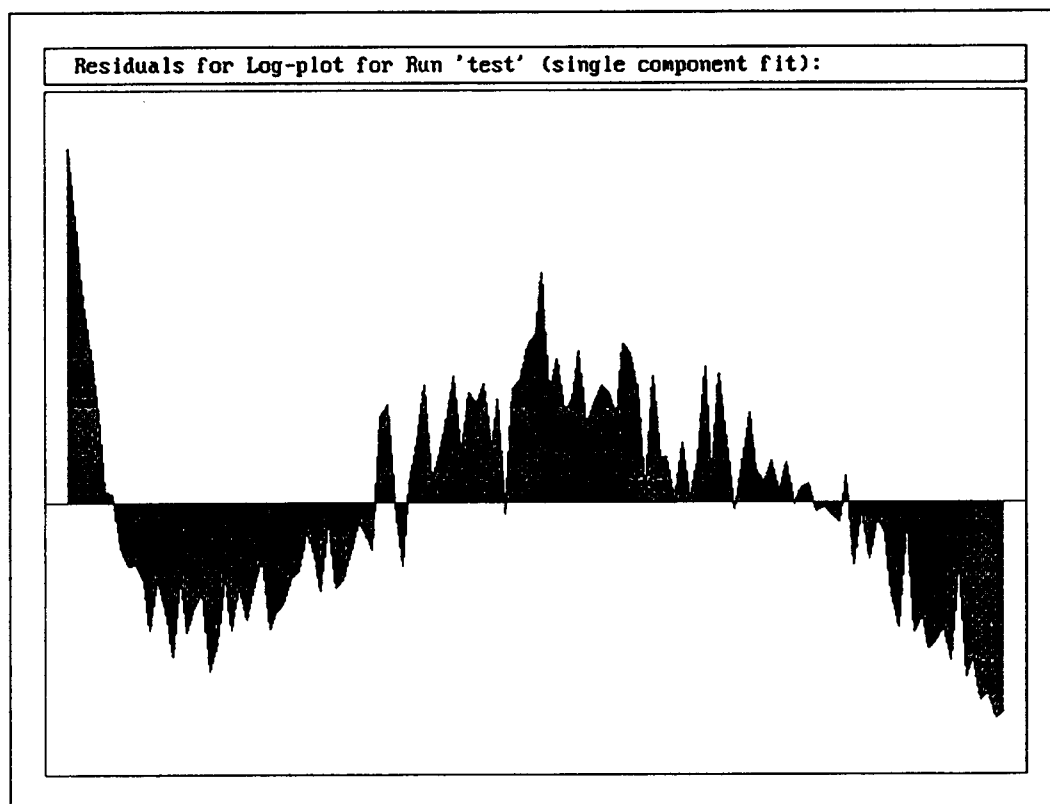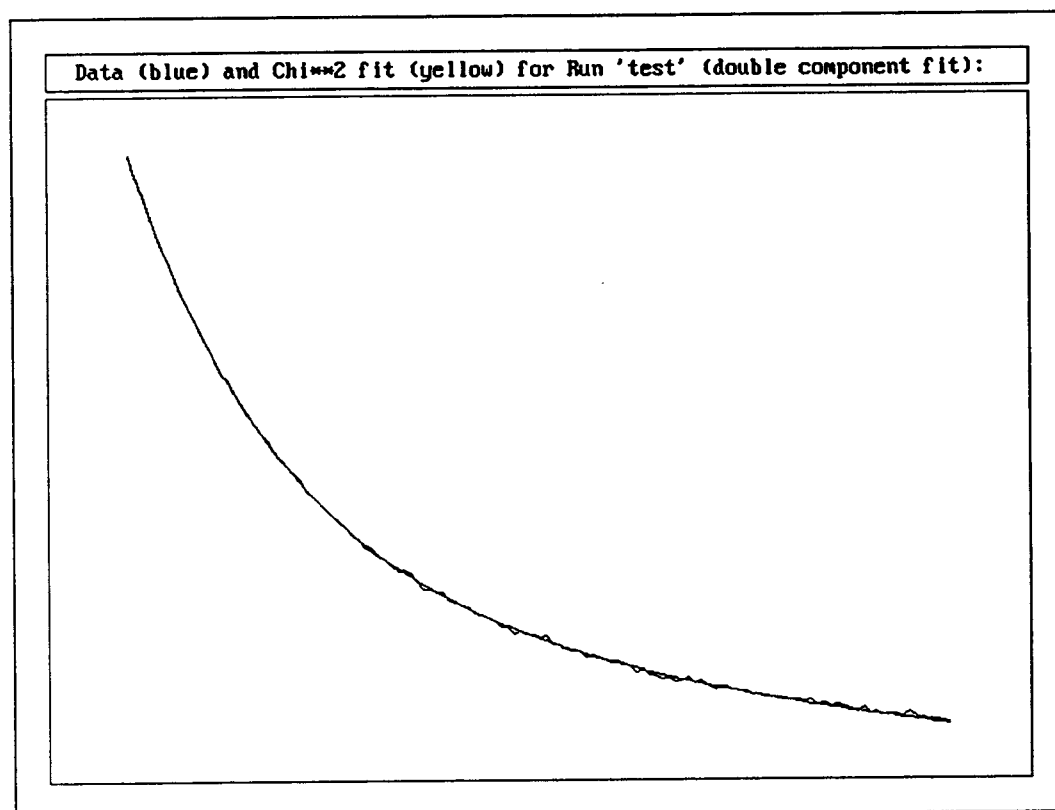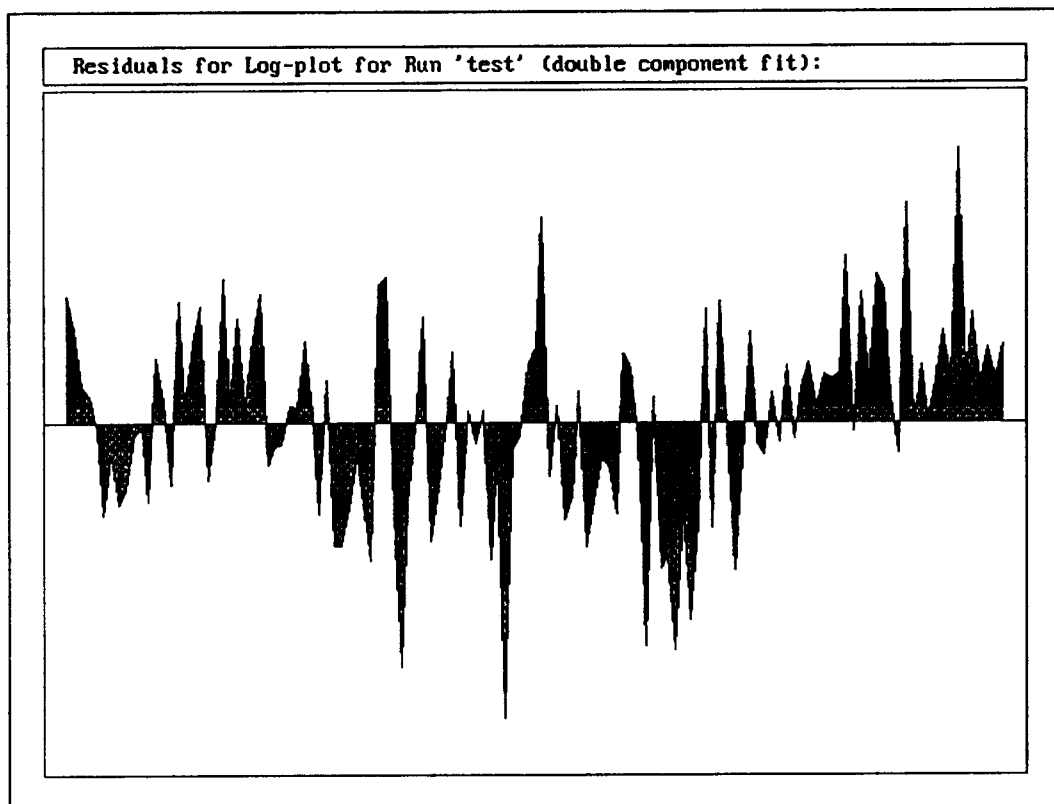
METAFILE PLOTS

The same plots can be obtained by introducing the metafile data into a spreadsheet plotting program like SUPERCALC, SIGMAPLOT, QUATTRO PRO etc. and plotting the data at higher resolution. To obtain higher resolution plots, use the meta files <run identification, sample time.1MT (single component) and .2MT (double component)>.

The data in these files is arranged in comma delimited columns for easy import into spreadsheet programs. The first column contains the sample time, the second column contains the actual data, the third column contains the data for the residuals plot and the fourth column contains the fitted data:

Sample time:	Actual data:	Residuals:	Fitted data:
.0000040,	11548211,	8726.0572,	11539484.9428
.0000080,	11539018,	13428.5987,	11525589.4013
.0000120,	11523773,	11777.9528,	11511995.0472
.0000160,	11508508,	9813.2715,	11498694.7285
.0000200,	11487748,	2066.5347,	11485681.4653
.0000240,	11473510,	561.5537,	11472948.4463
.0000280,	11459853,	-636.0240,	11460489.0240
.0000320,	11443033,	-5263.7110,	11448296.7110
.0000360,	11434805,	-1560.1762,	11436365.1762
.0000400,	11425345,	656.7590,	11424688.2410
.0000440,	11417734,	4474.1244,	11413259.8756
.0000480,	11402799,	724.8045,	11402074.1955
.0000520,	11387303,	-3822.4577,	11391125.4577
.0000560,	11374818,	-5590.0577,	11380408.0577
.0000600,	11372041,	2124.4742,	11369916.5258
.0000640,	11353534,	-6111.5239,	11359645.5239
.0000680,	11347639,	-1950.8428,	11349589.8428
.0000720,	11330576,	-9168.3981,	11339744.3981
.0000760,	11327233,	-2871.2284,	11330104.2284
.0000800,	11314916,	-5748.4913,	11320664.4913
.0000840,	11308354,	-3066.4611,	11311420.4611
.0000880,	11300161,	-2206.5257,	11302367.5257
.0000920,	11287415,	-6086.1843,	11293501.1843
.0000960,	11278080,	-6737.0440,	11284817.0440
.0001000,	11273795,	-2515.8179,	11276310.8179
.0001040,	11265270,	-2708.3220,	11267978.3220
.0001080,	11256021,	-3794.4731,	11259815.4731
...	...	...	...

etc...

METAFILE PLOTS

These are samples for the metafile plots (all plots have been prepared with SIGMAPLOT):

Run "TEST"

Sample time = $4.0 * 10^{-6}$, single component fit

Run "TEST"

Sample time = $4.0 * 10^{-6}$, residuals for single component fit

D-43
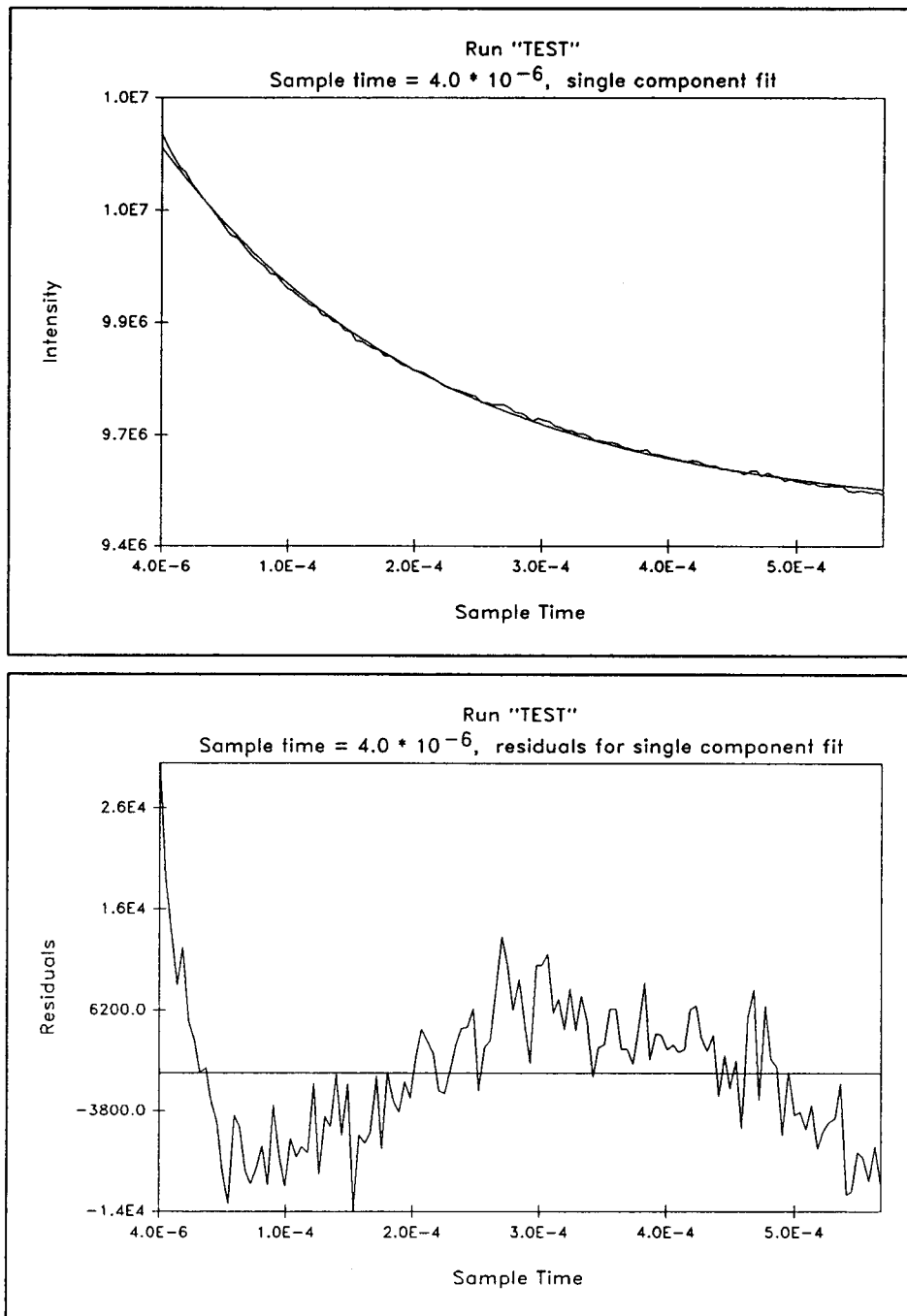
PRINT DATA RESULTS

The result files can be printed by selecting F2 from the print menu. First, the detailed report for each selected sample time is printed:

```
*********************************************************************
  RESULTS FOR FILE: TEST, Test run for Manual
*********************************************************************

Sample time:          4.0000e-06 sec
Temperature:          3.80 °C
Wavelength:           488.00 nm
Viscosity:            1.5769e+00 cp
Scattering angle:     90.00 degrees
Refractive index:     1.330
Total time of run:    3.8893e+02 sec
Sample time count:    9.7091e+07
Overflow count:       5.7510e+04
Shift reg. pulses:    3.5200e+08

  Laserscan Results and Statistics:
```

Data for single component fit: $A(\tau)=K1 \ast exp(K2 \ast \tau)+K3$

```
Diffusion Coeff.:     3.9238e-08 cm²/sec
Corrected Diff.:      6.5361e-08 cm²/sec

Amplitude (K1):       7.24126e+05    Stand. dev. σ(K1):   2.41851e-03
Frequency (K2):      -4.60242e+03    Stand. dev. σ(K2):   1.87381e-05
Baseline  (K3):       1.08225e+07    Stand. dev. σ(K3):   1.59648e-03
Estimated Baseline:   1.06625e+07

Statistical Tests:

F - Test:             0.00000e+00    Total Iterations:    2022
Chi - Square:         3.28644e+07
```

Data for double component fit: $A(\tau)=K1 \ast exp(K2 \ast \tau)+K3 \ast exp(K4 \ast \tau)+K5$

```
Diffusion Coefficient for 1. Component:           1.0422e-07 cm²/sec
Diffusion Coefficient, corrected to (20°,W):      1.7360e-07 cm²/sec
Relative Amount of 1. Component:                  52.37 %

Diffusion Coefficient for 2. Component:           1.5178e-08 cm²/sec
Diffusion Coefficient, corrected to (20°,W):      2.5284e-08 cm²/sec
Relative Amount of 2. Component:                  47.63 %

Amplitude-1 (K1):     5.18672e+05    Stand. dev. σ(K1):   4.35309e-03
Frequency-1 (K2):    -6.11215e+03    Stand. dev. σ(K2):   6.83966e-05
Amplitude-2 (K3):     4.71817e+05    Stand. dev. σ(K3):   1.42758e-03
Frequency-2 (K4):    -8.90175e+02    Stand. dev. σ(K4):   8.88770e-06
Baseline (K5):        1.05632e+07    Stand. dev. σ(K5):   2.89127e-03

Statistical Tests:

F - Test:             0.00000e+00    Total Iterations:    3002
Chi - Square:         1.72276e+07
```

After the individual results are printed, an abbreviated summary for all sample times is printed (file: <RUN_NUMBER>.LST):

```
                 Sample: Test, test for automatic run

*****************************************************************************
Corrected diffusion coefficients for sample time: .000003 sec

Single Component fit:      1.11850e-07 cm²/sec (100.00%) (X² = 1.93925e+07)
Double Component fit: (1)  1.92338e-07 cm²/sec ( 36.34%) (X² = 1.10219e+07)
                      (2)  3.08605e-08 cm²/sec ( 63.66%)
Weighted Average: (1 & 2)  8.95491e-08 cm²/sec
*****************************************************************************
Corrected diffusion coefficients for sample time: .000004 sec

Single Component fit:      1.16725e-07 cm²/sec (100.00%) (X² = 1.22928e+07)
Double Component fit: (1)  1.64248e-07 cm²/sec ( 44.64%) (X² = 9.45813e+06)
                      (2)  2.37469e-08 cm²/sec ( 55.36%)
Weighted Average: (1 & 2)  8.64682e-08 cm²/sec
*****************************************************************************
Corrected diffusion coefficients for sample time: .000005 sec

Single Component fit:      1.11624e-07 cm²/sec (100.00%) (X² = 1.44354e+07)
Double Component fit: (1)  1.59624e-07 cm²/sec ( 48.07%) (X² = 8.93450e+06)
                      (2)  2.23040e-08 cm²/sec ( 51.93%)
Weighted Average: (1 & 2)  8.83071e-08 cm²/sec
*****************************************************************************
Corrected diffusion coefficients for sample time: .000006 sec

Single Component fit:      1.02681e-07 cm²/sec (100.00%) (X² = 2.11341e+07)
Double Component fit: (1)  1.49167e-07 cm²/sec ( 49.49%) (X² = 1.16878e+07)
                      (2)  2.01315e-08 cm²/sec ( 50.51%)
Weighted Average: (1 & 2)  8.39857e-08 cm²/sec
*****************************************************************************
Corrected diffusion coefficients for sample time: .000007 sec

Single Component fit:      9.76237e-08 cm²/sec (100.00%) (X² = 2.16104e+07)
Double Component fit: (1)  1.41439e-07 cm²/sec ( 52.49%) (X² = 9.91794e+06)
                      (2)  1.87361e-08 cm²/sec ( 47.51%)
Weighted Average: (1 & 2)  8.31484e-08 cm²/sec
*****************************************************************************
Corrected diffusion coefficients for sample time: .000008 sec

Single Component fit:      9.31604e-08 cm²/sec (100.00%) (X² = 2.92283e+07)
Double Component fit: (1)  1.34575e-07 cm²/sec ( 54.84%) (X² = 1.42146e+07)
                      (2)  1.71805e-08 cm²/sec ( 45.16%)
Weighted Average: (1 & 2)  8.15545e-08 cm²/sec
*****************************************************************************
Corrected diffusion coefficients for sample time: .000009 sec

Single Component fit:      9.16029e-08 cm²/sec (100.00%) (X² = 3.10642e+07)
Double Component fit: (1)  1.32368e-07 cm²/sec ( 58.35%) (X² = 1.15140e+07)
                      (2)  1.76685e-08 cm²/sec ( 41.65%)
Weighted Average: (1 & 2)  8.45983e-08 cm²/sec
*****************************************************************************

etc...
```
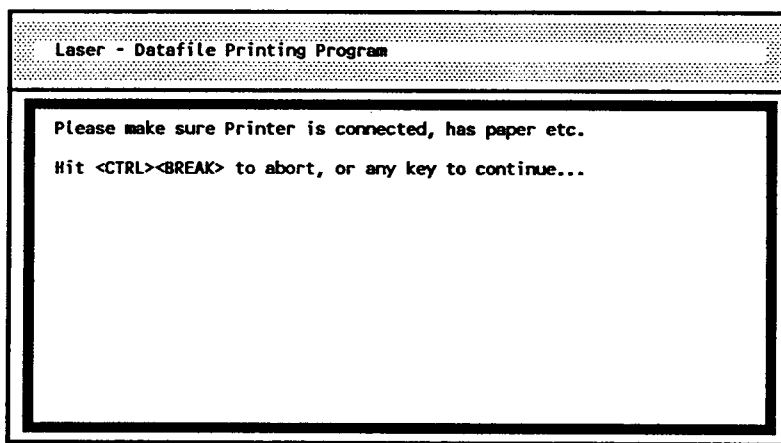
PRINTING ERRORS

If there is a problem with the printer, an error message will be displayed. This message can be due to several factors:

1. No paper in printer
2. No connection to printer (cable is loose, etc...)
3. Printer is not connected to LPT1 (parallel port). The printing output for the datafile is always directed to LPT1 as default. If your printer is connected to a different port, use the DOS MODE command to redirect the output to the desired port.

```
 Laser - Datafile Printing Program

  Please make sure Printer is connected, has paper etc.

  Hit <CTRL><BREAK> to abort, or any key to continue...
```

After the problem is fixed, retry the printing program. If everything is in order, following message will be displayed:

```
 Laser - Datafile Printing Program

  Printing in progress, please wait...
```
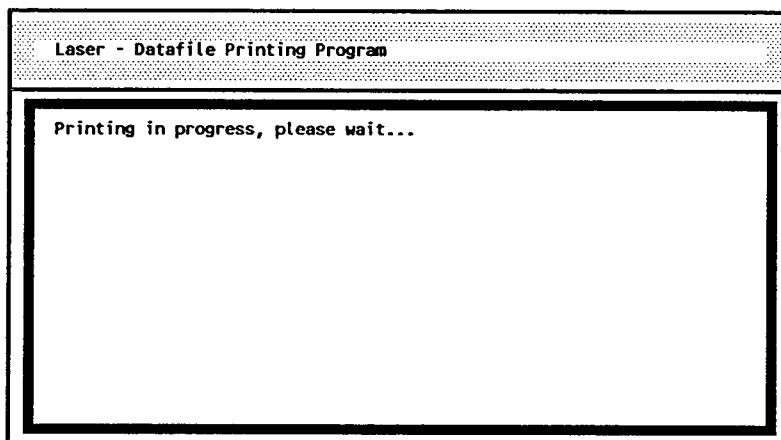
CHI-SQUARE PLOTTING PROGRAM

While the Levenberg-Marquardt χ^2 minimization routine provides a robust optimization algorithm for nonlinear models, disadvantages exist. The same χ^2 value can be obtained by different combinations of parameters. A χ^2 three dimensional surface and a χ^2 topographic surface can be constructed by simultaneously varying two of the obtained parameters by several standard deviations while keeping all others constant, and calculating the resulting χ^2 value for a two dimensional grid. The resulting χ^2 surface can then be used to visualize the confidence limits for each parameter with respect to another:
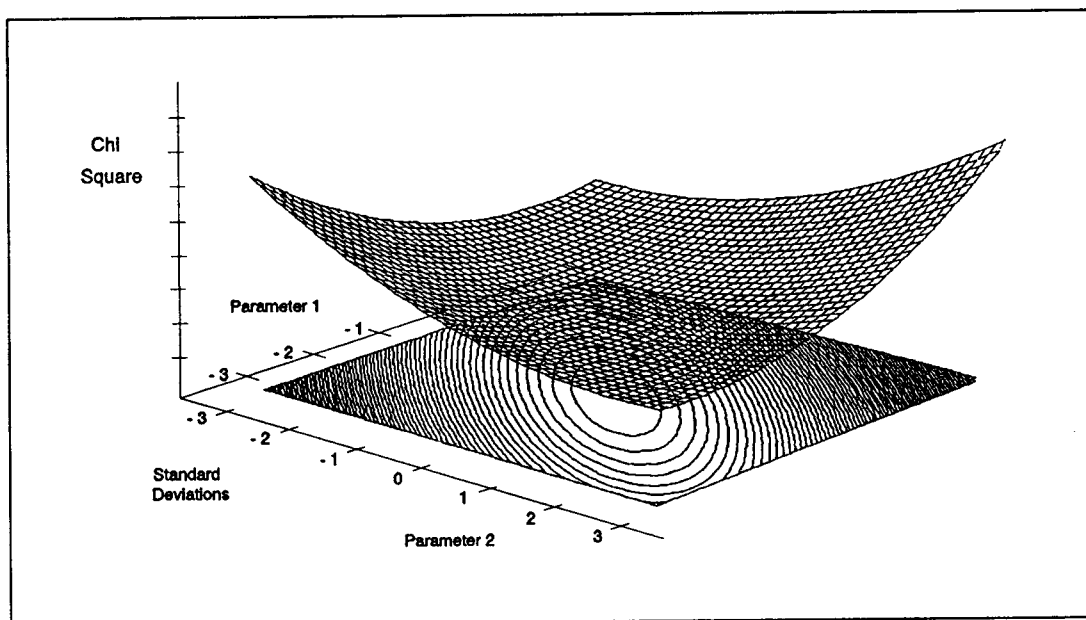
The plotting routine is invoked by selecting (F8) from the MAIN MENU. After selecting the run parameters, a selection menu is presented. Two parameters at a time can be checked against each other. Select any of the following:

D-47

```
┌─────────────────────────────────────────────────────────────┐
│   Sample: test, Automatic test run,          Sample time: 3.0*E-6   │
├─────────────────────────────────────────────────────────────┤
│  ┌───────────────────────────────────────────────────┐  │
│  │         Select Chi-square surface parameters:        │  │
│  │                                                       │  │
│  │   Parameters A1 and A2 (single component)            │  │
│  │   Parameters A1 and A3 (single component)            │  │
│  │   Parameters A2 and A3 (single component)            │  │
│  │   Parameters A1 and A2 (double component)            │  │
│  │   Parameters A1 and A3 (double component)            │  │
│  │   Parameters A1 and A4 (double component)            │  │
│  │   Parameters A1 and A5 (double component)            │  │
│  │   Parameters A2 and A3 (double component)            │  │
│  │   Parameters A2 and A4 (double component)            │  │
│  │   Parameters A2 and A5 (double component)            │  │
│  │   Parameters A3 and A4 (double component)            │  │
│  │   Parameters A3 and A5 (double component)            │  │
│  │   Parameters A4 and A5 (double component)            │  │
│  │            Exit Sample Time 3.0*E-6                   │  │
│  └───────────────────────────────────────────────────┘  │
└─────────────────────────────────────────────────────────────┘
```

After parameters for the plotting are selected, the program will prompt for the accuracy of the 3-dimensional grid and the number of standard deviations each parameter should be allowed to vary:

```
┌─────────────────────────────────────────────────────────────┐
│   Sample: TEST, Automatic test run           Sample time: 3.0*E-6   │
├─────────────────────────────────────────────────────────────┤
│                                                             │
│  Please enter a value for the grid density (suggested: 20): 20  │
│                                                             │
│  How many standard deviations for parammeter 1? (suggested: 10): 2  │
│                                                             │
│  How many standard deviations for parammeter 2? (suggested: 10): 4  │
│                                                             │
│                                                             │
│                                                             │
│                                                             │
│                                                             │
└─────────────────────────────────────────────────────────────┘
```
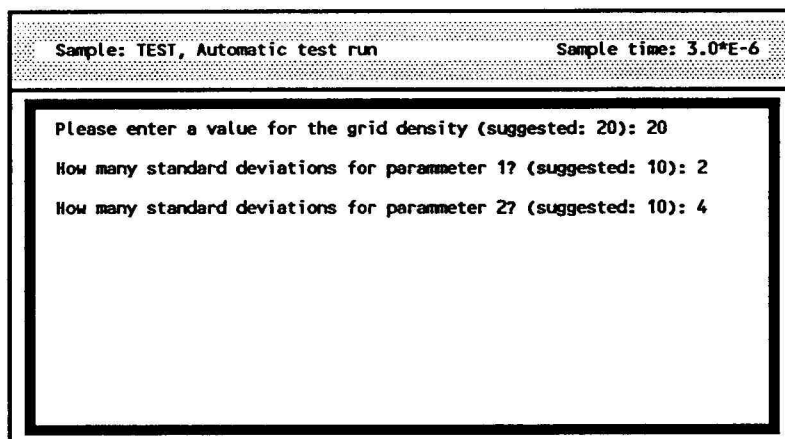
Even though a different standard deviation for each parameter can be selected, the scale for all parameters is always adjusted to be the same, creating always a square image. After selecting the grid size and the standard deviations, a 3-dimensional grid is calculated for the plot:

```
┌─────────────────────────────────────────────────────────────┐
│  Please wait, calculating chi-square surface for following parameters:  │
│  a1-single component and a2-single component. This might take a little while..  │
└─────────────────────────────────────────────────────────────┘
```

The first plot displayed is a topographical map showing lines af identical chi-square. The typically elliptical lines represent combinations of the parameters resulting in the same chi-square value:
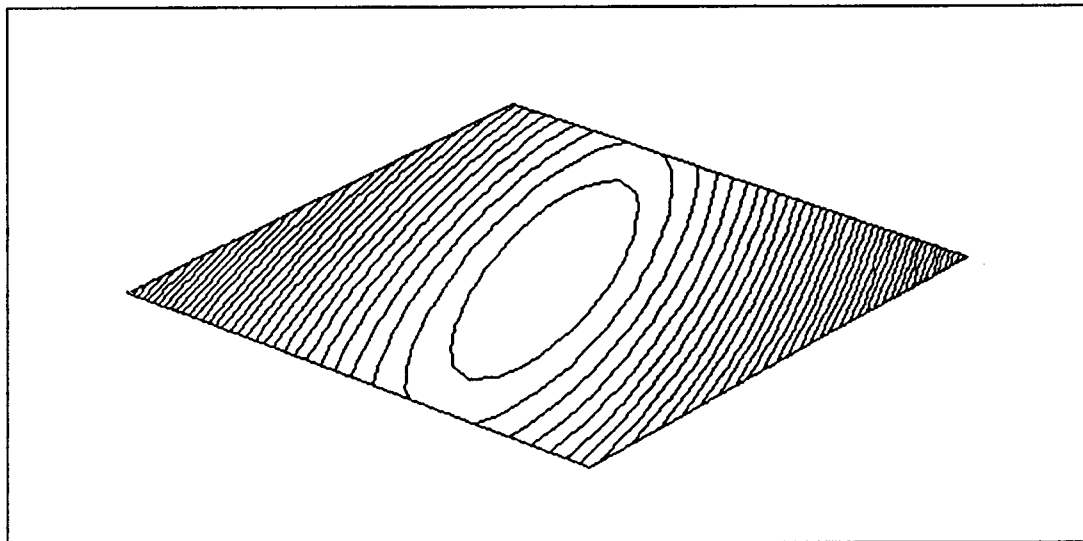
The second plot shows the 3-dimensional chi-square surface:
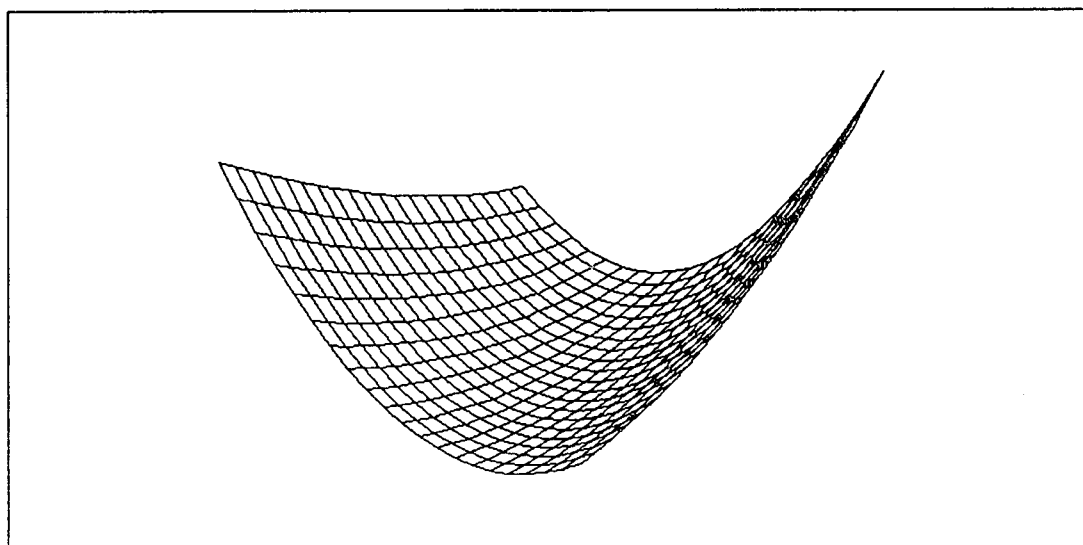
In the last plot, both graphs are displayed together:
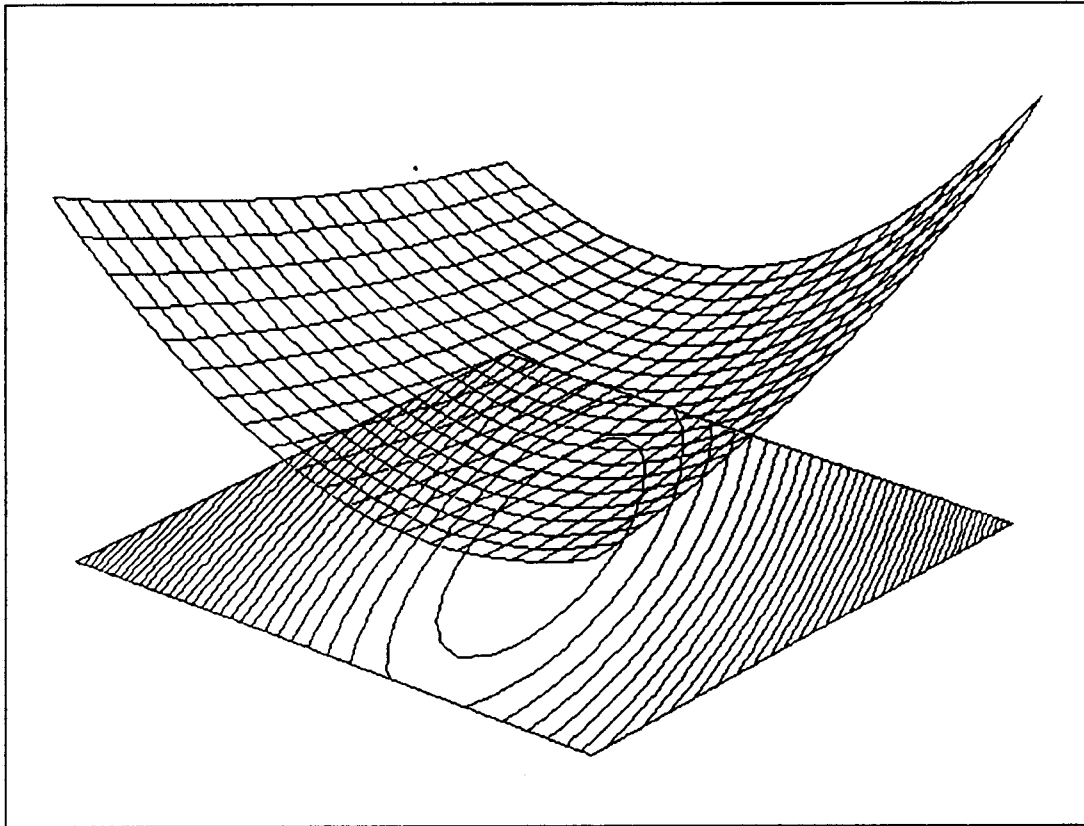
APPENDIX E

Neural Network Optimization for E. Coli Promoter Prediction

Borries Demeler and Guangwen Zhou

Department of Biochemistry and Biophysics

Oregon State University

Corvallis, OR 97331

Key Words: Promoter, Neural Network, Prediction, DNA sequence analysis.

ABSTRACT

Methods for optimizing the prediction of *Escherichia coli* RNA polymerase promoter sequences by neural networks are presented. A neural network was trained on a set of 80 known promoter sequences combined with different numbers of random sequences. The conserved -10 region and -35 region of the promoter sequences and a combination of these regions were used in three independent training sets. The prediction accuracy of the resulting weight matrix was tested against a separate set of 30 known promoter sequences and 1500 random sequences. The effects of the network's topology, the extent of training, the number of random sequences in the training set and the effects of different data representations were examined and optimized. Accuracies of 100% on the promoter test set and 98.4% on the random test set were achieved with the optimal parameters.

INTRODUCTION

An important objective in molecular biology is to analyze DNA sequences for structural and functional motifs. Efforts have been made to develop statistical algorithms for sequence analysis and motif prediction by searching for homologous regions or by comparing sequence information to a consensus sequence (1). These algorithms may fail or yield insufficiently accurate results when consensus sequences are difficult to define (1,2). In this paper we discuss the analysis of nucleic acid sequences by optimizing the parameters of a neural network implementation. The ability of optimized multilayered neural networks to predict structural and functional elements in DNA is demonstrated on bacterial RNA polymerase promoter sequences from *Escherichia coli (E. coli)*.

Neural networks have previously been applied to biological problems: The utility of neural networks for protein structure prediction from amino acid sequence data is documented in references (3-7). Brunak *et al.* (8) report the ability of a neural network to detect errors in the assignment of mRNA splice sites in nucleic acid sequence databases. While a perceptron model for the classification of *E. coli* translational initiation sites has been investigated by Stormo *et al.* (9), and a neural network model for the prediction of promoter sequences has been proposed by Lukashin *et al.* (10), a comprehensive analysis of neural network implementations for nucleic acid sequence analysis has not been presented so far. In this work we present an empirical approach to the optimal encoding of neural networks for promoter recognition with respect to parameters outlined in the abstract.

Neural networks were originally developed to simulate the learning process in the brain by modeling the brain's anatomy of neural cells connected by axons over different synaptic strengths. In a neural network, neural cells are replaced by interconnected computational units, called neurons. Axons are represented by symbolic connections between neurons, and synaptic strengths are represented by weights and thresholds (neuron biases), applied to the connections and neurons, respectively. During the training stage, a training set of inputs and corresponding targets is presented to the neural network. The initially random weights and thresholds are continuously modified according to a steepest gradient descent

learning algorithm. With each iteration, the learning algorithm attempts to minimize an objective function, generally the error function of the training set, which describes the difference between the observed output and the desired output. In its learned stage, the weights constitute the networks ability to recognize a pattern intrinsic to the training set. We believe that the ability of a neural network to accomplish pattern recognition tasks in an automated fashion is well suited for sequence analysis.

The bacterial promoter sequences chosen for this research contain two conserved regions, the -10 and the -35 region, centered about 10 bp (basepairs) and 35 bp upstream from the transcriptional start site. The conserved -10 and -35 regions are connected by a spacer region, which can be categorized into 3 classes, depending on the number of bases contained in the spacer region (2). A previous attempt to predict promoter sequences relies on six empirically developed tests to filter out false positives predicted by a consensus sequence match algorithm, which also determines the spacer class the predicted sequence belongs to. However, overall prediction accuracy reaches only 77% (1). While Lukashin *et al.* (10) report a maximal overall prediction accuracy of 93%, our optimized neural network implementation significantly improves on the prediction accuracy to 98.4% and allows sequence analysis of regions with poorly definable consensus sequence. Also, our training set does not require the introduction of gaps for perfect alignment, because a continuous sequence is used for the training. This allows analysis of sequences where alignment is difficult or impossible to define, a question not addressed by Lukashin *et al.* (10).

METHODS

Data: The promoter sequences were taken from the compilation of *E. coli* RNA polymerase promoter sequences by Hawley and McClure (11). Of those, 80 bacterial and phage promoters were used for the training database. The remaining 30 plasmid and transposon promoter sequences and promoters generated by mutation were used in a test set to evaluate our method. The promoter sequences were arranged into three independent training and corresponding test sets:

(A) 20 bases centered around the -10 region, which includes the TATAAT consensus sequence with the first T at the 12th position;

(B) 20 bases centered around the -35 region, which includes the TTGACA consensus sequence with the first T at the 10th position.

(C) 44 bases aligned as in (A) and containing the -35 region without a gap between the conserved -10 and -35 region.

A pseudo-random number generating program was used to generate random sequences with equal composition A,G,T,C, which were combined with the promoter sequences in the training set in a ratio of 1:1 - 20:1. The sequences were arranged by having 1 - 20 random sequences follow each promoter sequence in the training set, depending on the ratio of random vs. promoter sequences tested. Different random sequences were tried to determine if exchanging one random sequence for another could alter the result. Each test set for the random predictions contained 1500 random sequences in a format corresponding to (A), (B) or (C).

Calculations and network formulation:

The calculations were performed on a SUN 3/260 and on a 386 AT equipped with a math-coprocessor. The program was written in FORTRAN and was of our own design. It featured a feed-forward type network (Fig.1) using the delta rule in backpropagation and a momentum term as described in Wasserman (12). The input data was coded as a binary, orthonormal set of dimension 4, called CODE-4 (0001=C, 0010=G, 0100=A, 1000=T), or a linearly dependent set of dimension 2, called CODE-2

(00=A, 01=T, 10=G, 11=C). If CODE-4 was used, the number of input units was 4 times the number

of bases analyzed, similarly, in CODE-2 there were twice as many input units as bases in the test

sequence. The target to each input sequence was coded as '1' for promoters, and '0' for random

sequences. The forward pass was calculated by summing the products of all inputs to a neuron in the

hidden layer with their respective weights (weight$_{1,jk}$), adjusting with a neuron bias and operating on

the sum with an activation function F. The outputs of the activation function from the hidden layer

neurons were multiplied with their respective weights (weight$_{2,k}$) and served as inputs into the single

neuron of the output layer. For the kth neuron in the hidden layer:

$$X_{1,k} = \sum_{i=1}^{n} (weight_{1,ik} \cdot input_i) \qquad (1)$$

$$F(X_{1,k}) = \frac{1}{1 + e^{(-X_{1,k} + t_{1,k})}} \qquad (2)$$

where $t_{1,k}$ is the threshold adjustment for the kth neuron in the hidden layer. The output was determined

by summing $F(X_k)$ over all k and calculating the activation function F:

$$X_2 = \sum_{i=1}^{k} weight_{2,i} \cdot F(X_{1,i}) \qquad (3)$$

$$output = F(X_2) = \frac{1}{1 + e^{(-X_2 + t_2)}} \qquad (4)$$

The backpropagation algorithm used is outlined in reference (12), with the modification that the training

rate was continuously adjusted during training as a function of the derivative of the error function. This

measure helps to avoid network paralysis in a local minimum of the parameter hyperspace. If the

training process is trapped in a local minimum, the total error is larger than 0 and the derivative of the

error function approaches 0 asymptotically. In such a case increasing the training rate increases the

stepsize allowed in the gradient descent which aids in overcoming the hills of the errorsurface which

surround the local minimum. The error function E was evaluated according to:

$$E_i = \sum_{j=1}^{240} (target_j - output_j)^2 \qquad (5)$$

where i is the iteration and j is the number of the sequence in the training set. The training rate was adjusted according to

$$training\ rate = \frac{k}{(E_{i-1} - E_i)} \qquad (6)$$

where k is a constant, generally set to 10^{-6} and i is the iteration. The thresholds were adjusted by backpropagation like the weights, except that their inputs were always taken as 1. The weights were initially assigned random values in the range of 0.01 to -0.01.

The network was trained until the error level reached 10^{-4}. To determine the effect of the extent of training, weight matrices with error levels between 10 - 0.001 were used to test the prediction accuracy (training set C). Topologies ranging from 1 to 10 neurons in the hidden layer were tested in their prediction accuracy.

Evaluation of results: To evaluate our network, the trained network was used to predict both the random and promoter sequences in the test set. The output generated was in the range between 0 and 1 and in the format of real numbers. This allowed differentiation between strong and weak predictions. An output larger than 0.5 was interpreted as "promoter sequence", while outputs less than 0.5 were interpreted as "random sequences". Prediction accuracy was measured as the fraction of sequences in the test set predicted correct. Predictions were performed independently on promoter sets and random sets. The overall prediction accuracy of the network is given by:

$$(Fc_{random} * Fc_{promoter}) * 100\% \qquad (7)$$

where Fc_{random} denotes the fraction of correct predictions from the random test set and $Fc_{promoter}$ the fraction of correct predictions from the promoter set.

RESULTS:

Data representation: Initially, training set (A) was used to establish the better of the two encoding schemes. Each base in the training set was translated into binary code according to CODE-2 and CODE-4 and trained to an error level of 10^{-4}. The trained weight matrices were used to predict the test set of (A). At the same time, the number of neurons in the hidden layer were varied to determine the effect of network complexity. The results are summarized in Fig.2. A substantial difference in performance was noticed between CODE-2 and CODE-4.

Network topology: Network topologies of hidden layer size 1 - 10 neurons were tested. No clear indication for the optimal hidden layer size emerged. Modification of other parameters at the same time did not change this result. Therefore, we restricted our analysis to topologies involving 1 to 10 hidden layer neurons (the case of 0 neurons (perceptron) in the hidden layer is basically equivalent to the case of 1 neuron in the hidden layer, where the output neuron merely serves to add another neuron bias, so that the neuron biases for both neurons are shared).

Training set composition: To determine the optimal ratio of random sequences : promoter sequences (r:p ratio) we varied the composition of the training set with r:p ratios between 1:1 to 20:1. For training sets containing 20 bases the r:p ratio was found to be optimal at 2:1. Increasing the r:p ratio from 1:1 to 2:1 added an average of 5.8% accuracy over the different architectures examined. However, changing to a ratio of 3:1, the overall prediction accuracy dropped slightly by an average of 1% (Fig.3,A). For the training sets containing 44 bases, the optimal r:p ratio was found to be 5:1 (Fig.3,B), although results from training sets containing a ratio of 10:1 were very similar (0.15% difference). A larger increase in prediction accuracy was noticed in changing from 2:1 to 4:1 for (C) (0.9% difference). Changing to a r:p ratio of 20:1 decreased the promoter sequence prediction accuracy sharply, however, the random sequence prediction accuracy increased slightly (Fig.3,B). To determine if prediction results are sensitive to an exchange of one random sequence for another, different sets of random sequences were incorporated in the training set. No significant changes in the results for above parameters were observed

(data not shown).

Extent of training: Training was performed until the total network error on the training set reached 10^{-4}. Training was interrupted at intermediate stages (starting at error levels around 10) and the weight matrices analyzed (Fig.4). Generally, once training became asymptotic, prediction accuracy did not improve substantially. Maximum prediction accuracy was usually attained between error levels of 1.0 and 0.0001. The accuracy for promoter, random and combined predictions from 6 different training sets for (C) at various error levels is shown in Table 1.

Training set results: After establishing these parameters, prediction accuracy of all three training sets with their optimal parameters were compared (Fig.5). The optimal parameters (the parameters which give the highest prediction accuracy) of the three different training sets are listed in Table 2. Prediction accuracy of the training sets was found to be best for (C) with 98.4%, then (A) with 90% and (B) only 68.4%.

Consensus sequence prediction: Using the weight matrix obtained from the network with one neuron in the hidden layer (this case is equivalent to the perceptron formulation), it is possible to predict a consensus sequence. The value of the weights is plotted in groups of four, each element of a group representing one of four nucleotides in that sequence position. The value of the weight can be negative or positive, depending on whether the corresponding neuron is inhibitory or excitatory. The consensus sequence is determined by taking the most positive element of a group and relating it to the corresponding nucleotide. Strongly negative weights indicate nucleotides not favored in a particular position. Using such a diagram it is possible not only to extract a consensus sequence for a promoter, but also to determine the significance of each nucleotide in a particular sequence position. The result for a network with a single neuron for training set (C) with a r:p ratio 5:1 is shown in Fig.6.

Genomic sequence predictions: Promoter positions in phage fd genomic sequence (6408 basepairs) were predicted with our neural network. The prediction results are shown in Table 3 for a weight matrix with one hidden layer neuron from training set C (with phage fd promoters excluded). No significant

differences were found when other network architectures were employed. Promoters with prediction values between 0.5 and 0.9 were classified as weak promoters, those above 0.9 as strong promoters.

Of the eleven promoters determined by Schaller et al. (13), promoters fd V and fd I' were not predicted by the neural network, nine others were identified as strong promoters (r:p ratios between 2:1 and 10:1). Overprediction decreased for matrices from 2:1 to 10:1. Although overprediction was further decreased by changing from an r:p ratio of 10:1 to 20:1, only seven promoters were correctly determined with an r:p ratio of 20:1. Statistical methods only predicted five promoters correctly (14). Obviously, the major problem in the prediction of promoters is overprediction. Increased r:p ratios will, in general, decrease overprediction, but also decrease promoter prediction accuracy, a result which agrees with our artificial sequence predictions. The results from artificial sequences suggest that overprediction between 2% - 4% is to be expected. For phage fd, 120 - 240 mispredicted promoters would be expected. Preliminary results using a larger database (including additional promoters from Harley and Reynolds (15)) indicate that overprediction can be reduced significantly (data not shown).

DISCUSSION:

From the results summarized in Fig.2, it seems that CODE-4 is the better choice for DNA data representation. This is a unitary coding matrix with identical Hamming distance among each vector. Unitary coding has been used previously for encoding amino acids in secondary protein structure prediction by neural networks (5,7) and perceptron analysis of nucleic acids (8,9). CODE-2 representation features linearly dependent vectors, where the Hamming distance among vectors is different (Hamming distance is a measure for the difference between vectors by counting the number of different bit entries, i.e., a vector 1,1 has a Hamming distance of 1 compared to the vector 0,1 while the Hamming distance is 2 compared to the vector 0,0. In CODE-4 encoding all vectors have an equal Hamming distance of 2 between each other. Only comparisons between vectors of the same dimension is meaningful). This presumably could bias the learning process towards certain bases. Although we did not test the effects of CODE-2 vs. CODE-4 on training set (B) and (C), we expect the results to be similar. Combined prediction accuracy increased by an average of 9% if CODE-4 was used. While Lukashin et al. (10) used CODE-2 encoding, Brunak et al. (8) and Stormo et al. (9) employed CODE-4 encoding for their studies. We checked the effect of different hidden layer size simultaneously with each parameter modified. Hidden layer size did not seem to influence the results obtained in a predictable fashion and differences were only minor. While the representational power of the network increases with added neurons, the generalization ability of the network decreases. If the amount of data increases (by using larger r:p ratios) the need for larger networks arises. This was exemplified by the failure of smaller networks to train below a certain error level. If the representational power was increased by adding neurons and weights, this problem was usually overcome. Training behavior was also strongly dependent on the initial, randomized weight matrix used. Failure to train could be overcome by modifying the initial weight matrix, thus providing the network with a starting point closer to the learned state.

The r:p ratio in training set (A) and (B) seems to be optimal at 2:1, although we did not test ratios higher than 3:1. Since the number of possible sequences is much lower for training sets containing 20

bases (1.1 * 10^{12}) than for training sets containing 44 bases (3.1 * 10^{26}), it seems not surprising that a higher r:p ratio results in better predictions for the 44 base training set.

As Holley and Karplus report in (5), we also observe that training beyond a certain error level (10^{-4}) does not add to the prediction accuracy. This might be due to loss of generalization through learning of the idiosyncracies of the sequences in the training set. Training to a level where the network error is below 1.0 means that all sequences in the training set can be correctly classified as "random" and "promoter". The optimal extent of training for r:p ratios of 4:1 and 3:1 were obtained at an error of 1.0 and 0.1, respectively. When the r:p ratio is increased to 5:1, the best prediction is obtained at an error of 0.0001. This might be due to the additional information provided to the network with the additional random sequences in the training set. While promoter predictions generally improve with more extensive training, random prediction accuracy could be improved only by using a higher r:p ratio. There is a trade-off, which needs to be optimized, as promoter prediction accuracy naturally decreases with added random sequences (Fig.3,B).

In comparison of (A), (B) and (C), significant differences between predictions of the -10 region and the -35 region were observed. The differences between (A) and (B) are presumably due to a greater correlation between the promoter sequences in the -10 set. This suggests that the -35 region is less conserved than the -10 region of E. coli promoter sequences. An average difference of 23% was observed, with (A) giving the better result. Although the prediction accuracy of training set (B) is substantially lower (68.4%) than that of (A) (90%), it is still above 50%, which is the accuracy expected if no classification can be made at all. This suggests that the informational content of (B) could add to that of (A), if (A) and (B) were combined to (C). Here it is important to note that an alignment question arises: The sequences of (A) are separated by three different spacing classes from the sequences of (B) (2). We aligned the combined sequences both as in (A) and (B) in independent training sets. The results indicated that alignment of the TATAAT box regions in (A) is preferential to alignments of the TTGACA consensus sequence in (B) (data not shown). Combination of (A) and (B) as in (C) provided

an additional prediction accuracy of 3.5%.

We believe that neural networks can provide a valuable alternative method in the analysis of DNA or RNA sequences, as demonstrated by the prediction of promoter sequences. In addition to attaining a prediction accuracy of 98.4%, our findings show that:

1. CODE-4 encoding for nucleic acid data should be used in preference to CODE-2 encoding.

2. The number of hidden layer units does not have a significant effect on prediction accuracy.

3. A low network error level improves promoter prediction accuracy, while a larger r:p ratio improves the network's ability to filter out false positives. For promoter predictions, we found the optimal r:p ratio to be between 5:1 and 10:1.

4. A combination of the conserved regions around -10 and -35 into a sequence of 44 bases length aligned at the -10 region yields better prediction results than a sequence of 20 bases length around either conserved region alone.

5. The utility of our method is characterized best by the improvement in prediction accuracy over previous methods. While the promoter prediction accuracy is high, the overall prediction error is caused mainly by overprediction, i.e. false positives. Therefore, we believe that our method is useful for the localization of possible promoter sequences, which could be further verified by biochemical means.

We suggest that the overall prediction accuracy could be increased by including more promoter sequences in the training database. We plan to investigate effects of various AT:GC ratios in random training sequences on the prediction accuracy. Since promoter sequences contain a higher ratio of AT:GC than 1:1 (as in our random sequences), inclusion of random sequences with higher AT:GC ratios in the training database may lead to a more sensitive classification between promoters and non-promoters.

ACKNOWLEDGMENTS:

REFERENCES:

1. O'Neill, M.C. and Chiafari, F. (1989) J. Biol. Chem. **264**, 5531-5534.

2. O'Neill, M.C. (1989) J Biol. Chem. **264**, 5522-5530.

3. Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B. and Petersen, S.B. (1990) FEBS Letters. **261**, 43-46.

4. Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Noerskov, L., Olsen, Q.H. and Petersen, S.B. (1988) FEBS Letters. **241**, 223-228

5. Holley, L.H. and Karplus, M. (1989) Proc. Natl. Acad. Sci. USA. **86**, 152-156.

6. Kinoshita, J. (1990) Sci. Amer. **262**, 24-26.

7. Qian, N. and Sejnowski, T.J. (1988) J. Mol. Biol. **202**, 865-884.

8. Brunak, S., Engelbrecht, J. and Knudsen, S. (1990) Nuc. Acids res. **18**, 4797-4801

9. Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982) Nuc. Acids Res. **10**, 2997-3011

10. Lukashin, A.V., Anshelevich, V.V., Amirikyan, B.R., Gragerov, A.I. and Frank-Kamenetski, M.D. (1989) J. Biomol. Struct. & Dyn. **6**, 1123-1133

11. Hawley, D.K. and McClure, W.R. (1983) Nuc. Acids Res. **11**, 2237-2255.

12. Wasserman, P. D. (1989) Neural Computing, Theory and Practice. (Van Nostrand Reinhold, New York, NY). p. 6-59.

13. Schaller, H. and Beck, E. and Takanami, M. (1978) In Denhardt, D.T., Dressler, D. and Ray, D.S. (ed.) Single Stranded DNA Phages. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y. pp 139-163.

14. Mulligan, M.E. and McClure, W.R. (1986) Nucl. Acids Res. **14**, 109-126.

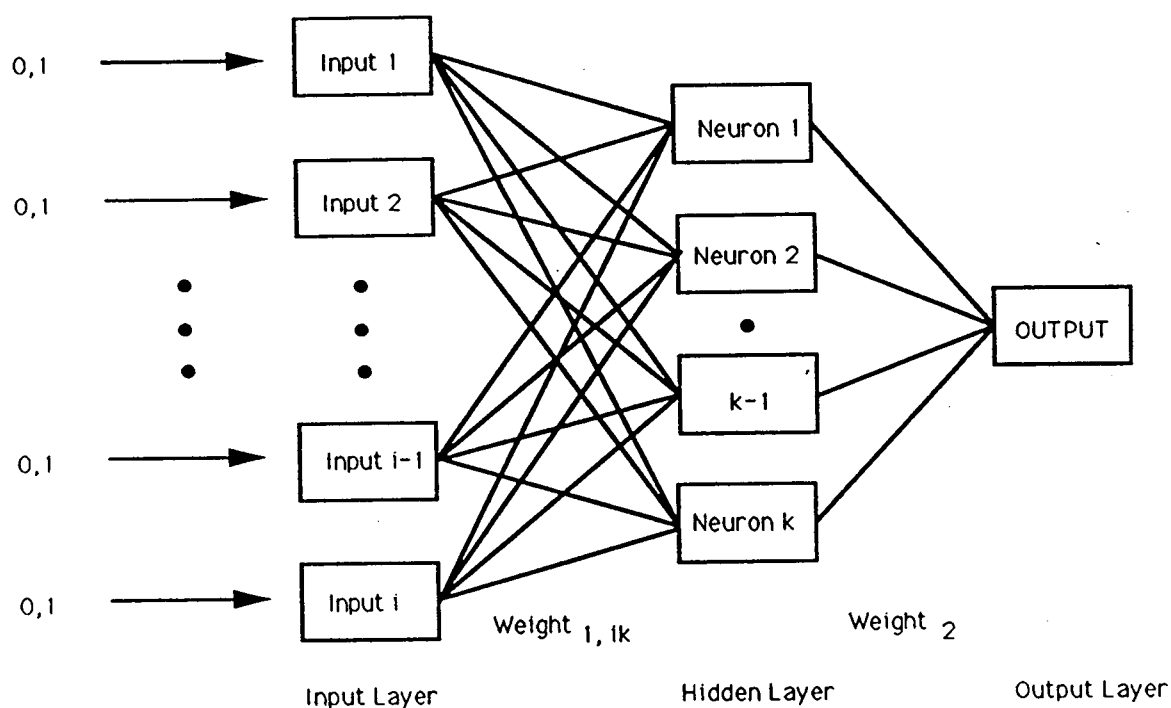15. Harley, B.C. and Reynolds, R.P. (1987) Nucl. Acids Res. **15**,

Figure 1:
Shown is a neural network architecture with 4 neurons in the hidden layer, k=4. The number of input units is either 4 times the number of bases (CODE-4) or 2 times the number of bases in the input sequence (CODE-2). The values of the input units can be 0 or 1. There is one output unit and it is in the range between 0 and 1. Weights$_1$ are applied to the connections between the input and the hidden layer, weights$_2$ are applied to the connections between the hidden and output layer. Only the neurons in the hidden and output layer serve a computing function, the inputs only serve to distribute.
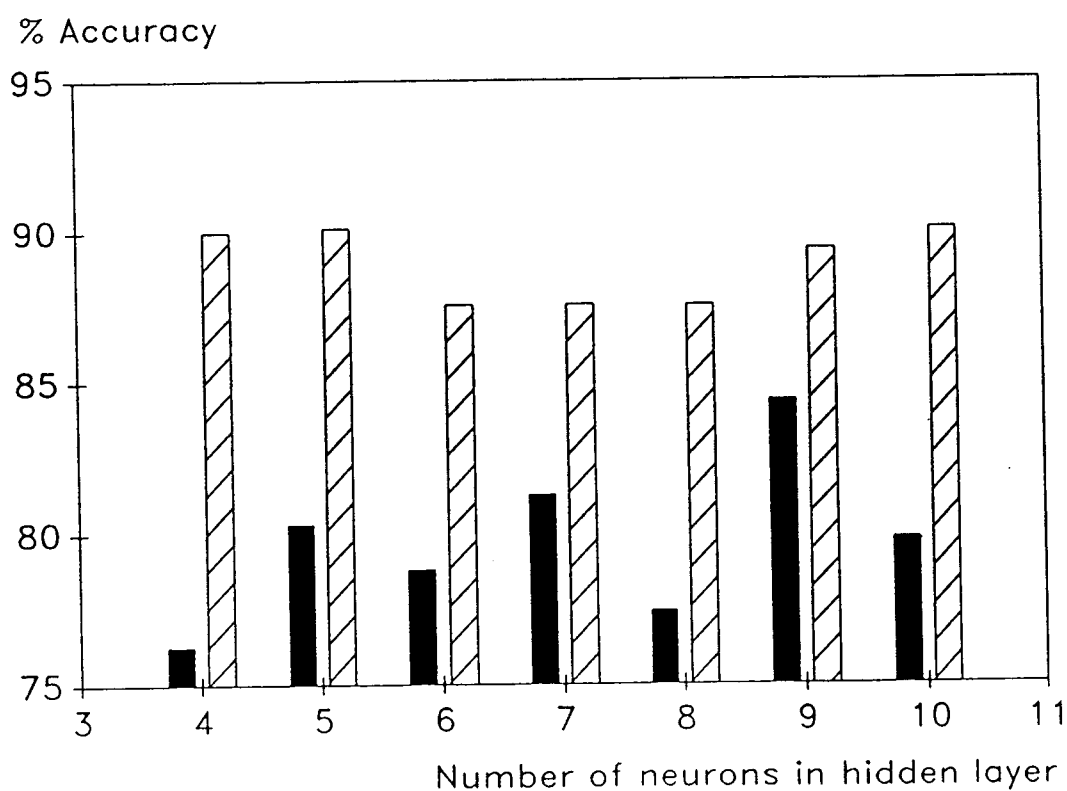
Figure 2:
Training results comparing CODE-2 with CODE-4.
The %-accuracy for the combined predictions for CODE-2 is shown in black bars, prediction results for CODE-4 is shown in hatched bars. Network topologies ranging between 4 and 10 hidden layer neurons are compared, when the error level was trained to 10^{-4}.
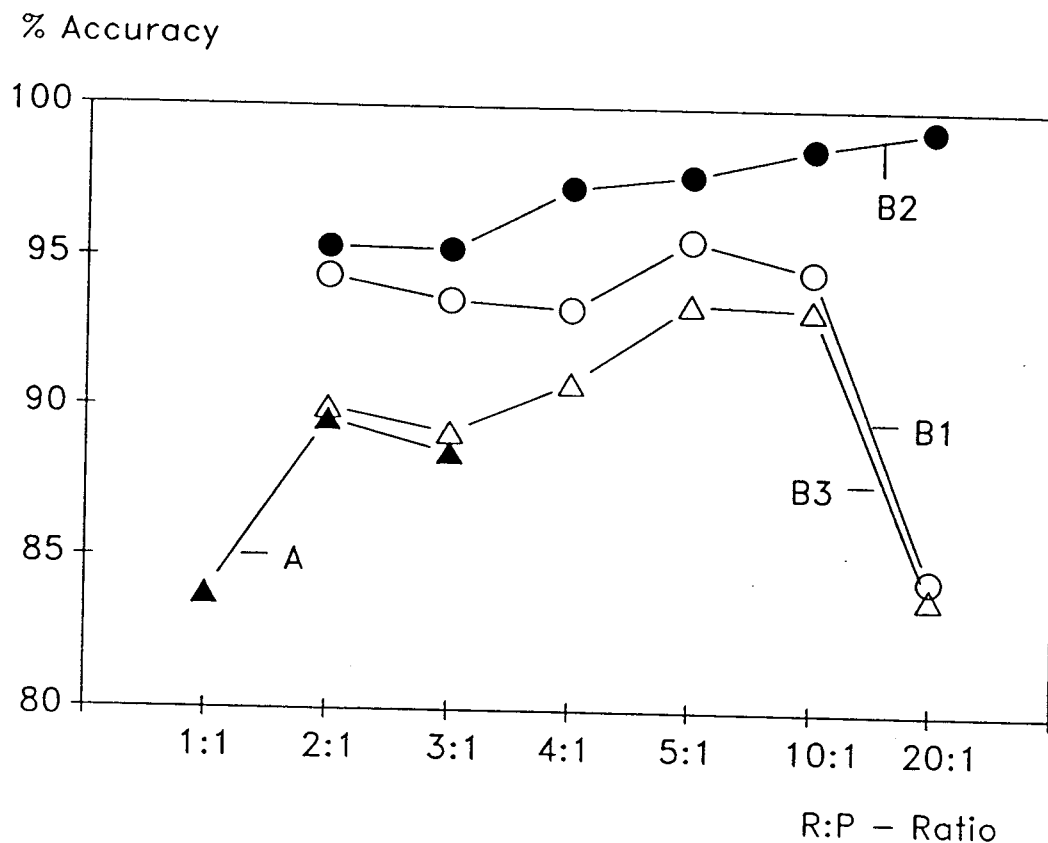
Figure 3:
(A) Accuracy of combined prediction results for r:p ratios ranging between 1:1 and 3:1 for 20 bases in the input sequence. The data was obtained by averaging results from network topologies ranging between 4 and 10 neurons in the hidden layer.
(B) Accuracy of combined prediction results for r:p ratios ranging between 2:1 and 20:1 for 44 bases in the input sequence.
(B1): promoter predictions
(B2): random predictions
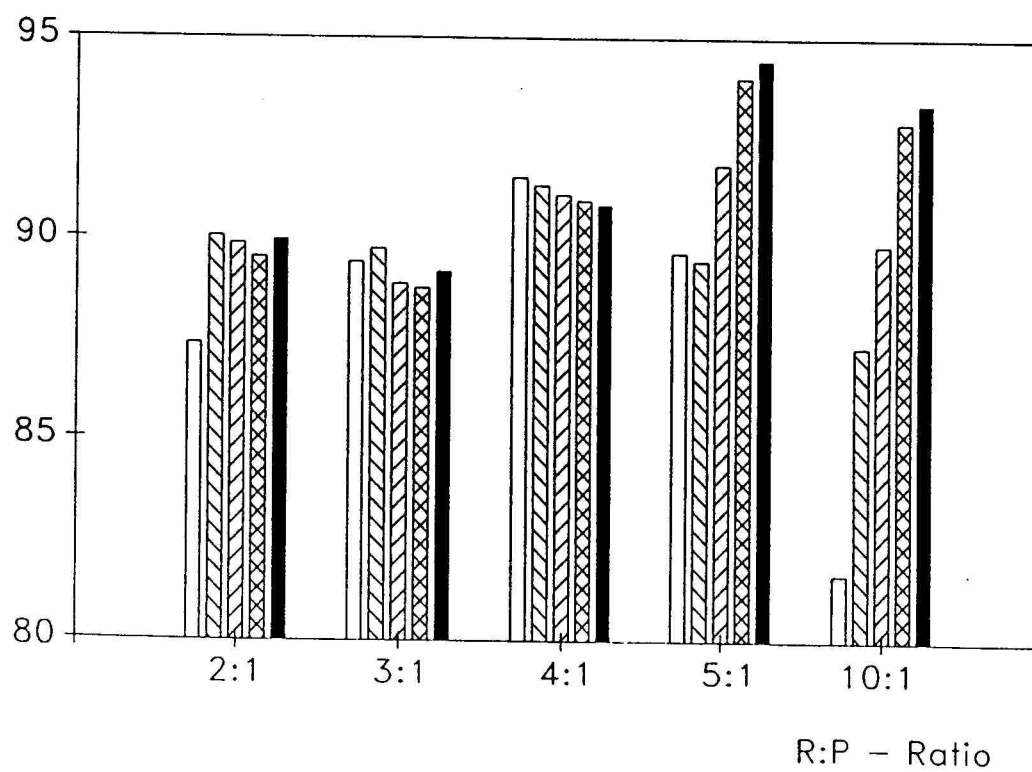(B3): combined predictions

% Accuracy

Figure 4:
The effect of varying the extent of training for training set (C) with a r:p ratio between 2:1 - 10:1.
Results shown are averaged from network topologies between 1 and 10. From left (blank bar) to right
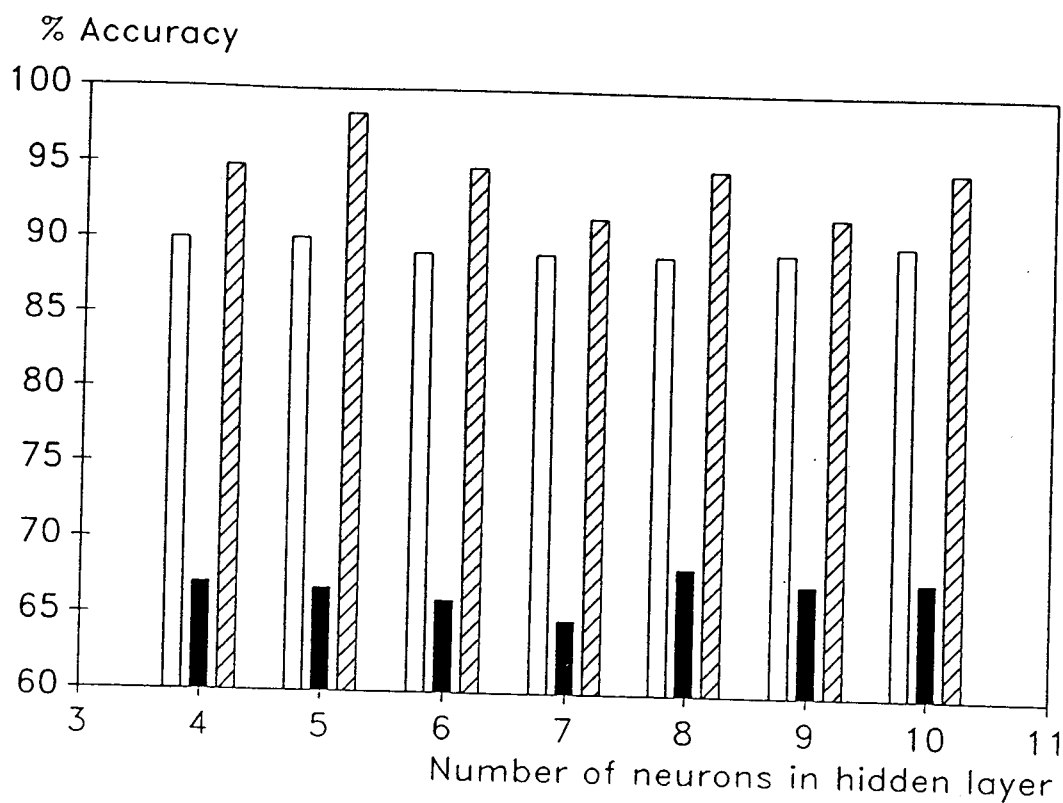(filled bar) error level = 1, 0.1, 0.01, 0.001, 0.0001.

Figure 5:
Combined prediction results for the three different training sets. Results are shown for different network topologies, ranging from 4 to 10 neurons in hidden layer.
(A): training = 10^{-4}, r:p = 2:1 (white bar)
(B): training = 10^{-4}, r:p = 2:1 (black bar)
(C): training = 10^{-4}, r:p = 5:1 (hatched bar)

Table 1:

Effect of Extent of Training on Prediction Accuracy:

ERROR LEVEL:	r:p	1.0	0.1	0.01	0.001	0.0001
PROMOTER PREDICTIONS:	2:1	90.29	93.57	93.82	93.73	94.37
	3:1	93.59	93.59	93.10	93.10	93.59
	4:1	93.33	93.33	93.33	93.33	93.33
	5:1	91.93	91.79	93.10	93.97	95.64
	10:1	82.26	88.17	90.86	94.06	94.62
	20:1	76.34	79.03	82.26	83.87	84.41
RANDOM PREDICTIONS:	2:1	96.74	96.24	95.79	95.52	95.35
	3:1	95.52	95.89	95.45	95.33	95.29
	4:1	98.07	97.86	97.61	97.46	97.35
	5:1	98.30	98.22	97.87	97.92	97.79
	10:1	99.26	99.01	98.89	98.78	98.70
	20:1	99.52	99.56	99.55	99.60	99.34
COMBINED PREDICTIONS:	2:1	87.35	90.05	89.87	89.53	89.98
	3:1	89.40	89.74	88.86	88.75	89.18
	4:1	91.53	91.33	91.10	90.96	90.86
	5:1	90.37	90.16	91.12	92.02	93.53
	10:1	81.65	87.30	89.85	92.91	93.39
	20:1	75.97	78.95	81.89	83.53	83.85

Shown are combined prediction results for training set (C). Combined prediction results were obtained by averaging results from all networks with 1–10 neurons in the hidden layer. Results for different r:p ratios are shown for different stages in the error level of the training process. The combined prediction accuracy is highest for a r:p ratio of 5:1 and an error level of 0.0001.

Table 2:

Optimal Parameters for Training Sets A, B and C:

training set:	r:p ratio:	encoding scheme:	extent of training:	maximal accuracy:	number of neurons
(A)	2:1	CODE-2	0.0001	88.85	9
(A)	2:1	CODE-4	0.0001	90.12	5
(B)	2:1	CODE-4	0.0001	68.39	8
(C)	5:1	CODE-4	0.0001	98.40	5

Shown are prediction results for the optimal configurations from each training set. The maximal accuracy is the combined accuracy of promoter and random predictions.

Table 3:

Prediction results for genomic sequence data of phage fd:

r:p	Error	1	0.1	0.01	0.001	0.0001
2:1	Weak	481	354	259	213	167
	Strong	153	441	567	637	687
4:1	Weak	387	257	184	142	105
	Strong	117	318	405	467	503
5:1	Weak	336	264	183	131	96
	Strong	135	329	434	487	523
10:1	Weak	224	179	122	96	77
	Strong	99	202	282	329	360
20:1	Weak	47	46	53	38	27
	Strong	152	179	179	141	150

Prediction results for phage fd. Promoters are classified as
"weak" (prediction between 0.5 and 0.9) and "strong" (>0.9).
More extensive training increases the number of strong
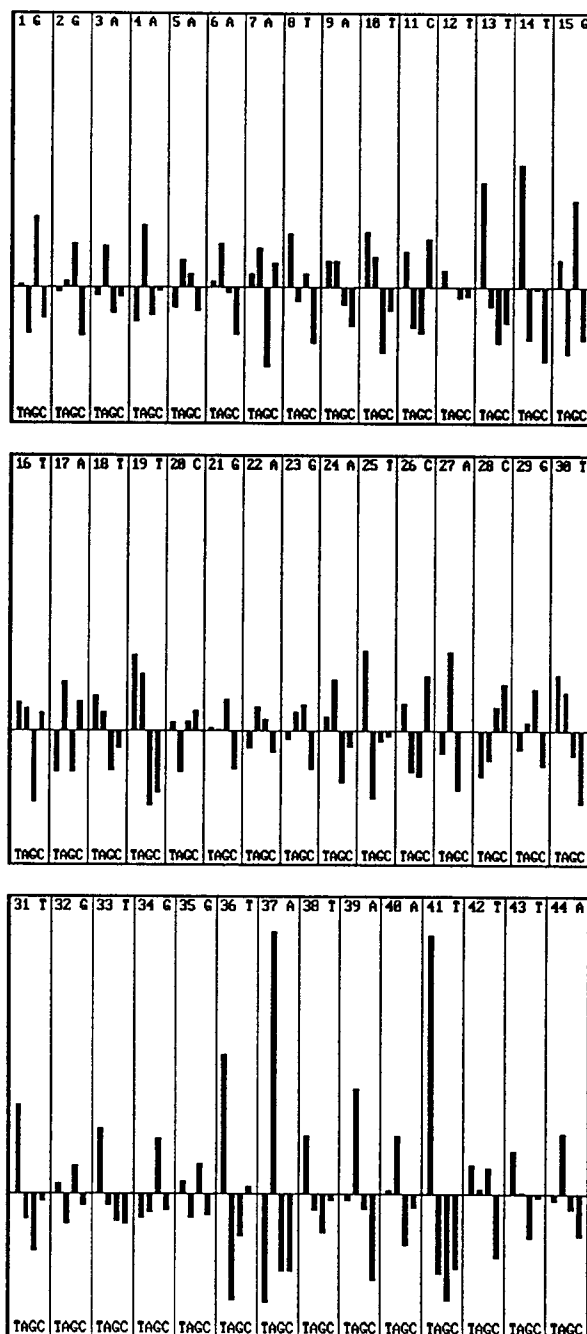promoter predictions, larger r:p ratios favor a decrease in
overprediction.

Figure 6:
Weight diagram for the single neuron case. The consensus sequence is indicated on top, together with the respective sequence position. Weights are shown as excitatory (above middle line) or inhibitory (below middle line). Each sequence position has a certain value for each possible nucleotide.