STUDIES ON THE ARRANGEMENT OF REPEATED SEQUENCES IN DNA

Thesis by

William Raymond Pearson

In Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1977

(Submitted February 24, 1977)

Acknowledgements

I would like to thank my advisor, Dr. James Bonner, for his guidance, patience and understanding during my graduate career. I also thank Dr. E. Davidson, for his invaluable discussions, comments and encouragement during this work. I appreciate the work of my other committee members, G. Attardi, N. Davidson and J. P. Revel.

Much of the work described in this Thesis was done in collaboration with others. Chapter I was done in collaboration with J. R. Wu with assistance from J. Posakony. J. R. Wu and J. Posakony also helped with Chapter II. Chapter IV was done with assistance from S. Smith and R. Wilson.

I am especially grateful for the discussions, encouragement and understanding of my good friends, Jung-Rung Wu and Margaret Chamberlin.

Abstract

Parameters of repetitive DNA sequence organization have been measured in the rat, Drosophila and the pea genomes.

Experiments using melting, hydroxyapatite binding and single strand nuclease digestion have been used to measure the number length and arrangement of repeated DNA sequences in the rat. About 20% of rat DNA is repeated 3000 fold. Half of the sequences are 200 - 400 nucleotides long while the remainder are longer than 1500 nucleotides. Rat repeated DNA sequences are interspersed among 2500 nucleotide long single copy sequences.

Studies on the long and short repeated DNA sequences of the rat show that the long repeated sequences are also 3000-fold repeated. Cross-hybridization of isolated long and short repeated sequences and hydroxyapatite binding interspersion experiments indicate long and short repeated DNA may share sequences, although this may be due to cross-contamination. Self-renaturation, melting and electron microscopy of long repeated DNA fragments suggest some long fragments may be composed of arrays of shorter repeated sequences.

A similar sensitive search has been made in Drosophila melanogaster for short repetitive sequences interspersed with single copy sequences. Five kinds of measurements all yield the conclusion that there are few short repetitive sequences in this genome: 1) Comparison of long and short fragment reassociation kinetics; 2) reassociation kinetics of long fragments driven by an excess of short fragments; 3) measurement

of the size of repeated fragments after S-1 nuclease digestion; 4) measurement of the hyperchromicity of repeat sequence bearing fragments of different lengths; 5) direct assay by kinetics of reassociation of the amount of single copy sequence present on 1200 nucleotide long fragments which also contain repetitive sequences.

Renaturation of pea DNA has been used to estimate the size of the pea genome and the fraction of pea DNA containing repeated DNA sequences. Pea DNA renaturation and single copy tracer hybridization indicate the size of the pea genome is 0.45 pg. More than 70% of pea DNA is repeated from 100 to 5000 times.

Table of Contents

Chapter I	Analysis of Rat Repetitive DNA Sequences	1
Chapter II	Analysis of Sequences in Rat Repetitive DNA A preliminary Report	55
Chapter III	Absence of Short Period Interspersion of Repetitive and Non-repetitive Sequences In the DNA of Drosophila melanogaster	106
Chapter IV	Kinetic Determination of the Genome Size of the Pea	125
Appendix A	A Program for Least Squares Analysis Of Reassociation and Hybridization Data	149

CHAPTER I

ANALYSIS OF RAT REPETITIVE DNA SEQUENCES

Introduction

Evidence is accumulating that one of the mechanisms for differential gene expression is the sequence specific regulation of RNA transcription. Recent work on a variety of organisms has demonstrated an almost universal highly ordered pattern of repetitive sequence organization in DNA (see Davidson et. al., 1975a for review). The proximity of subsets of repeated sequences to transcribed and translated DNA sequences supports a regulatory function. These structural observations on sequences near functional DNA sequences suggest that repeated sequences may play an important role in DNA sequence recognition during the regulation of transcription.

Repeated DNA sequences display strikingly similar, highly ordered structures across a wide range of organisms from insects to mammals (Davidson et. al., 1973a, 1973b; Bonner et. al., 1973; Graham et. al., 1974; Angerer et. al., 1975; Chamberlin et. al., 1975; Goldberg et. al., 1975; Schmid and Deininger, 1975; Efstratiadis et. al., 1976). With a few exceptions (Manning et. al., 1975; Crain et. al., 1976), repeated sequences 200 to 400 nucleotides long are interspersed among 1000 to 2000 nucleotide non-repeated single copy sequences in more than 65% of the DNA of all organisms studied. This organization was predicted by Britten and Davidson (1969; also Davidson and Britten, 1973) in a model which suggested that repeated DNA sequences can coordinately control the expression of adjacent single copy genes.

While the similarity of organizational patterns across the evolutionary spectrum is striking evidence for repeated DNA function in gene expression, experiments demonstrating specific subsets of repeated sequences adjacent to transcribed and translated genes provide even stronger support for the regulatory function of repeated sequences. Experiments on large nuclear transcripts in the rat (Holmes and Bonner, 1974b) and sea urchin (Smith et. al., 1974) show that most large nuclear RNA is transcribed from DNA containing interspersed repeated sequences. More recent experiments suggest that a select subset of repeated sequences are near transcribed and translated sequences. Gottesfeld has reported (1976) that a fraction of DNA from chromatin with increased transcriptional activity contains not only a subset of single copy DNA but also a subset of all repeated sequences. Experiments using sea urchin messenger RNA have shown that 80% of the translated sequences are adjacent to repeated DNA sequences (Davidson et. al., 1976b) and that these adjacent sequences are a subset of the repetitive sequence population (Klein et. al., in prep.). Selected repeated sequences are adjacent to single copy DNA sequences which are transcribed into message and translated.

In this paper we describe experiments which measure the structural parameters of repeated DNA sequence organization in a mammal, the rat. We have made three basic measurements. First, the fraction of the DNA which contains repeated sequences has been determined. Second, we have measured the length and thus the number of the repeated sequences. Third, we have studied the distribution of repeated sequences in single copy DNA.

Materials and methods

Preparation of DNA

DNA used in these experiments was prepared from two sources. Unlabeled DNA was isolated from a rat Novikoff ascites cell line grown intraperitoneally and transfered at seven day intervals. Cells removed from the rats were frozen and stored at -80 °C. For the DNA preparation, cells were thawed and a crude nuclear pellet isolated by the method of Dahmus and McConnell (1969). The crude nuclei were resuspended in 0.1 M Tris, 0.1 M EDTA pH 8.5 and lysed with the addition of 20% sodium dodecyl sulfate to a concentration of 2%. The viscous lysate was heated to 60 °C for 15 minutes and then digested with pronase (Calbiochem, grade B) at 50 ug/ml until the DNA went into solution (usually two hours). The DNA solution was then extracted with an equal volume of 1 : 1 phenol:IAC (24:1 chloroform:isoamyl alcohol), the aqueous phase removed and the interphase re-extracted with buffer and phenol:IAC. The aqueous phases were pooled and re-extracted with phenol:IAC followed by 2-3 extractions with IAC. The DNA was then precipitated with 2.5 volumes of 95% ethanol at room temperature and spooled. The DNA was dissolved in 10mM Tris 10 mM EDTA pH 8.5 overnight at 4 °C, then brought to 0.1 M Tris 0.1 M EDTA pH 8.5 and digested with 50 ug/ml RNase A (Worthington) for 90 minutes (after 10 min RNase preincubation at 95 °C) after which 200 ug/ml preincubated pronase (90 min, 37 °C) was added. After a second 90 minute incubation at 37 °C the DNA was extracted three times with IAC, precipitated with ethanol and spooled. The DNA was redissolved and spun at 27,000 RPM in a Beckman SW27 rotor to remove undissolved material, and then reprecipitated and

spooled.

^{3}H labeled DNA was prepared from Novikoff hepatoma cells grown in suspension culture in flasks on a shaker. The cells were grown on a modified Swimm's 67 medium (Plageman and Swimm, 1966; Plageman, personal communication) and had a doubling time of 12 hours. Cells were labeled with ^{3}H thymidine by the addition of 0.1 uCi/ml (Schwartz, 46 mCi/umole) four times at 12 hour intervals. Under these conditions, we were able to isolate DNA with specific activities around 100,000 cpm/ug.

After growth to a concentration of 1 x 10^{6} cells/ml the cells were centrifuged at 2000g for 2 min and resuspended in lysis buffer, 10 mM Tris, 10 mM NaCl, 1.5 mM $MgCl_2$, 1 mM $CaCl_2$, 1 mM triethanolamine pH 7.4 (B4; Plageman, 1969), pelleted at 2K for 5 min, resuspended in B4 and allowed to stand for 5 min. The cells were lysed with 10 strokes of a glass homogenizer and the nuclei spun down at 2000g for 5 minutes. This crude pellet was then extracted as described above for the unlabeled ascites DNA.

Preparation of DNA fragments

DNA fragments of various sizes were prepared by shearing in a Virtis 60 homogenizer (Britten et. al., 1974). 4000 nucleotide fragments were prepared by shearing at 7500 RPM and 800-1100 nucleotide fragments were sheared at 30,000 RPM for 45 min. 350 nucleotide DNA was prepared by shearing at 50,000 RPM in 66% glycerol as described by Britten et. al. (1974). Broader distributions of fragment lengths for interspersion studies were prepared by shearing at 2000 to 40,000 RPM from 30 sec to 5 min and individual sizes fractionated from preparative

alkaline sucrose gradients. After shearing, all preparations were filtered and passed over Chelex-100 (BioRad) and then precipitated from 0.3 M NaAcetate with 2.5 volumes of 95% ethanol.

Sizing DNA fragments

Single stranded DNA fragment lengths were determined by sedimentation through alkaline sucrose gradients. Isokinetic sucrose gradients (Noll, 1967) were formed in SW41 tubes in 0.1 N NaOH using a Vmix of 10.4 ml, Cflask = 16 % w/v, Cres = 43% w/v. Gradients were centrifuged from 16 to 24 hours at 40,000 RPM. All tubes contained two markers of known molecular weight and samples were run at least two times. Molecular weights were calculated from sedimentation rates using the Studier (1965) equations.

S-1 nuclease renaturation and digestion

DNA samples to be digested with S-1 nuclease were incubated in 0.3 M NaCl 0.01 M Pipes pH 6.8 at 65 ^{o}C (Smith et. al., 1975; Britten et. al., 1977). The equivalent Cot was calculated using a factor of 2.31 to correct for the 0.3 M Na^{+} concentration. Long (>500 nucleotide) DNA samples were alkaline denatured to ensure complete denaturation.

After incubation, samples to be nucleased were diluted with an equal volume of 0.05 M NaAcetate 0.2 mM $ZnSO_4$ pH 4.2 and dithiothreitol was added to 5mM. The final reaction mix was 0.05 M Pipes 0.025 M NaAcetate 5 mM pipes 0.1 mM $ZnSO_4$ 5mM DTT at pH 4.4. The pH was checked in each experiment. The S-1 nuclease preparation used was the gift of Dr. F. Eden and D. Painchaud and has been extensively characterized

(Britten et. al., 1977). The standard enzyme to DNA ratio used was 15 ul/mg (DIG = .78; Britten et. al., 1977) for measurements of the fraction repetitive in the genome. DNA samples were incubated with S-1 nuclease for 45 min at 37 °C. The reaction was terminated by addition of 2.0 M PB to a final concentration of 0.12 M and the sample passed over hydroxyapatite. For most experiments, the fraction bound to hydroxyapatite was denatured at 100 °C and eluted with 0.12 M PB. The DNA was not denatured before A-50 chromatography but eluted with 0.5 M PB. The size distribution of S-1 nucleased duplexes was measured on a 110x1.3 cm column of agarose A-50 (Bio-rad). The gel bed was poured around a support of 5 mm glass beads (Britten et. al., 1974). Samples were chromatographed in 0.12 M PB using $^{32}PO_4$ as an inclusion marker.

DNA renaturation

Samples which were not to be digested by S-1 nuclease were incubated in 0 .12 M PB at 60°C or in 0.48 M PB at 70 °C. After incubation, samples were frozen in dry ice-ethanol. Samples were thawed and diluted to 0.12 or 0.14 M PB and passed over hydroxyapatite at 60 °C. The fraction bound was eluted after thermal denaturation at 100 °C in all experiments except the melt presented in Figure 4. The fraction and rate parameters for the renaturation curves were calculated using a non-linear least squares fitting program (Pearson et. al., 1977a).

Melting

DNA samples were melted in 0.12 M PB in a Gilford model 2400 spectrophotometer equipped with a model 2527 thermal cuvette. Samples were melted at a rate of 0.5 ^{o}C/min and the A_{260} automatically sampled at 0.5 ^{o}C intervals. Hyperchromicity was calculated from the formula

$$H = \frac{A260(98^{o}C) - A260(60^{o}C)}{A260(98^{o}C)}$$

after subtraction of the buffer absorbance at each temperature.

Results

The renaturation of rat DNA

Repeated sequences can be characterized from the kinetics of renaturation of short DNA. Figure 1 shows the renaturation of 350 nucleotide long fragments. The fraction of the fragments containing duplex was determined by hydroxyapatite binding. Least squares computer fits to these data are summarized in Table I. The data display three qualitatively distinct components: a single slow component renaturing between Cot 200 and 20,000; a more rapidly renaturing component binding between Cot 0.02 and 200 and a very rapidly reannealing component bound by the earliest times in the data.

The slowly renaturing component containing 55-65% of the DNA is single copy non-repeated DNA. Table I shows two different fits of the data analysing this component. The first fit, a "free fit", allows all parameters to vary to find the best fit. This fit does not terminate but will continue beyond these parameters to a parameter set with more

Figure 1: Renaturation of 350 nucleotide rat DNA fragments

These data have been collected by a number of investigators (Holmes and Bonner, 1974a; Gottesfeld et. al., 1976). Denatured 350 nucleotide DNA was renatured in .12 M PB at 60 ^{o}C or in .48 M PB at 70 ^{o}C. The fraction single stranded was measured by hydroxyapatite chromatography. Equivalent Cot is the Cot (moles/liter-seconds) times a salt concentration factor (Britten et. al., 1974).

The solid lines show the least squares fit of the data using a single copy rate fixed at 0.00034 for a genome size of 2.9 pg (Sober, 1968). The actual coefficients for this fit are shown in Table IB.

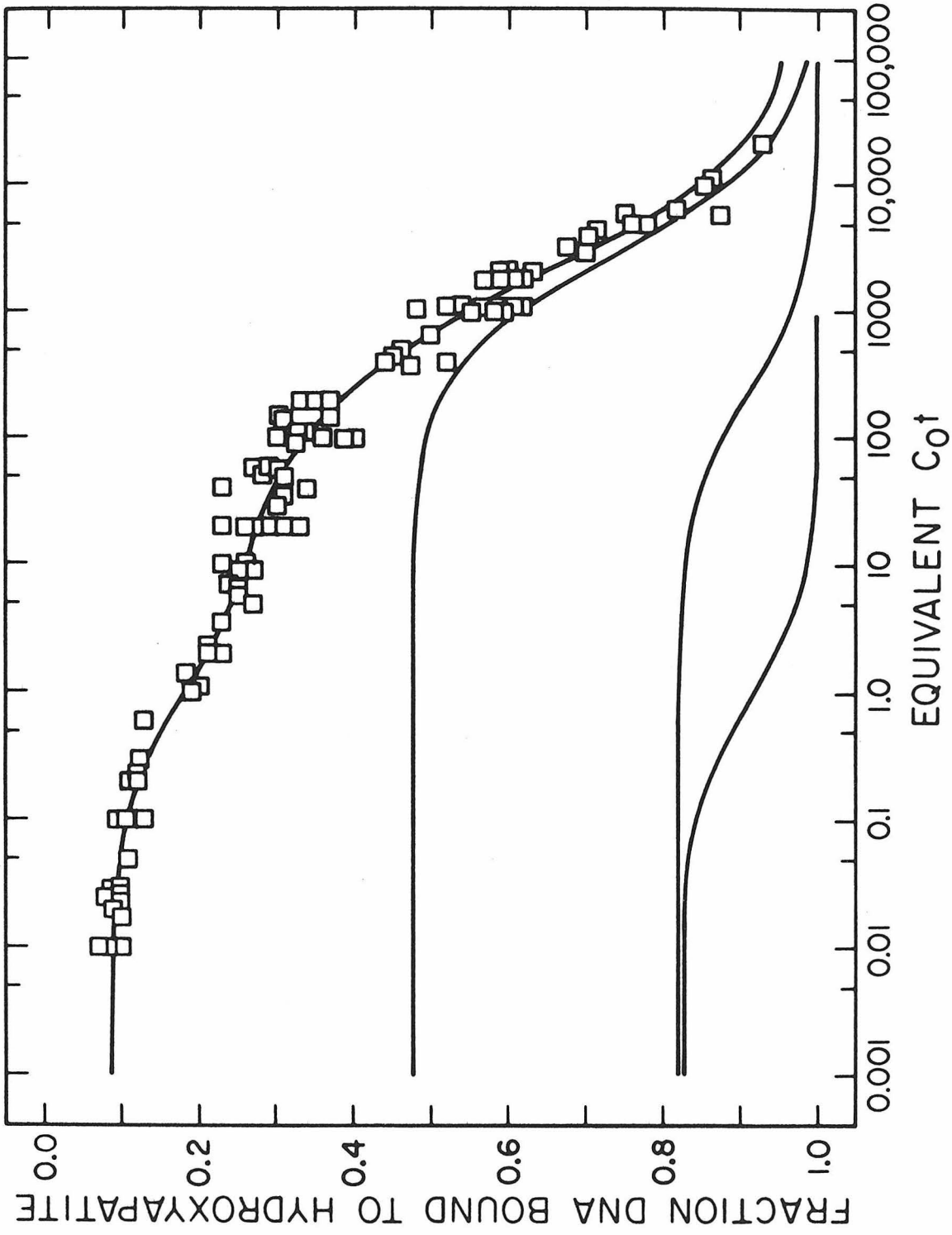

FRACTION DNA BOUND TO HYDROXYAPATITE
0.0
0.2
0.4
0.6
0.8
1.0
EQUIVALENT C_0t
0.001
0.01
0.1
1.0
10
100
1000
10,000
100,000

Table I

Renaturation of 350 nucleotide Rat DNA fragments

A. Unconstrained fit: (no parameters fixed)

Goodness of fit: 3.0%

Component	Fraction	Rate	Cot 1/2	Repetition Frequency
1	0.0883	> 100	0.01	>350,000
2	0.177 (±0.015)	1.06 (±0.39)	0.9434	3500
3	0.191 (±0.098)	0.00381 (±0.00308)	262.5	11
4	0.537 (±0.079)	0.000292 (±0.000132)	3424	1
	Final fraction unreacted:		0.0067 ±0.0453	

B. Constrained fit: (single copy rate fixed at 0.00034 for 2.9 pg genome)

1	0.0882	> 100	0.01	>350,000
2	0.176 (±0.014)	1.10 (±0.39)	0.909	3235
3	0.181 (±0.042)	0.00421 (±0.00211)	238	12
4	0.525 (±0.051)	0.00034	2941	1
	Final fraction unreacted:		0.0298 ±0.0194	

C. Constrained fit: Single copy and repetitive rates fixed

1	0.0729	> 100	0.01	>350,000
2	0.176 (±0.011)	2.97[a]	0.337	8727
3	0.213 (±0.020)	0.00412[a]	243	12
4	0.499 (±0.030)	0.00034	2941	1

Final fraction unreacted: 0.0391 ±0.0165

Footnotes to Table I

[a] Fixed repetitive rate constants calculated from the fit of the data in Figure 2.

than 100% reaction (Pearson et. al., 1977a). In the second fit, the rate constant for the single copy component was fixed at a number appropriate for the genome size of the rat. For a rat genome size of 2.9 picograms (Sober, 1968) the single copy rate constant is fixed at 0.00034 liter/mole-sec corresponding to a Cot1/2 of 2900. This fit terminates properly and we can calculate the fraction in the single copy component and the fractions and repetition frequencies of the repetitive components.

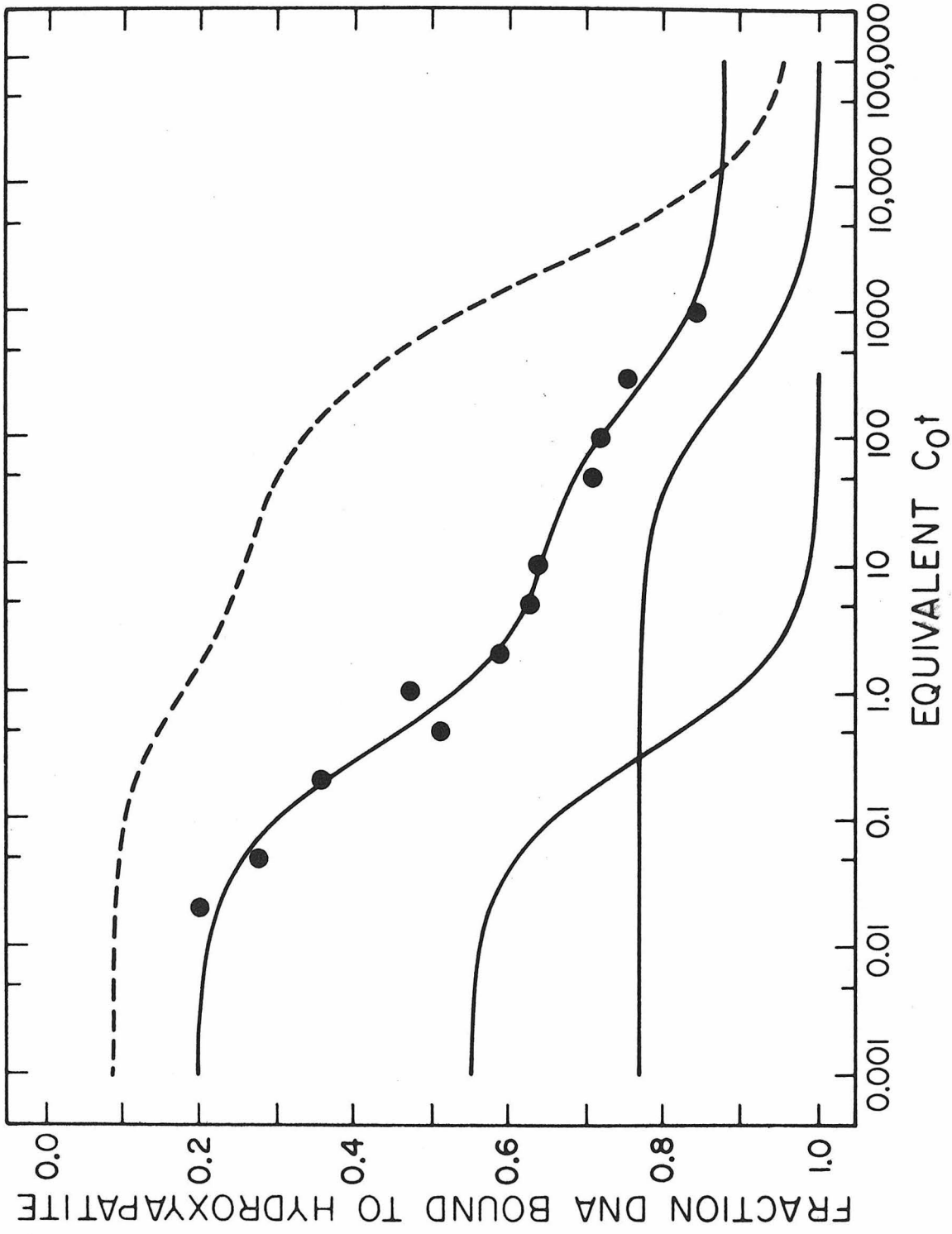
The renaturation from Cot 0.02 to 200 is due to the reaction of sequences repeated from 10 to 100,000 times in the rat genome. About 25% to 30% of the 350 nucleotide long fragments contain duplex in this region of the curve. The DNA fragments bound to hydroxyapatite contain single strand tails, and the real fraction of repeated sequences in the genome is lower. Due to the scatter in the data and the fraction of the genome involved, accurate estimation of the repetition frequencies for the different components is difficult. This is reflected in the high parameter error estimates.

For better resolution of the repetitive components a repetitive fraction of the genome was isolated. ^{3}H labeled 350 nucleotide long fragments were renatured to Cot 100 and the duplexed repetitive sequences separated on hydroxyapatite. This repetitive fraction was then hybridized with a 100 - 1000 fold excess of whole 350 nucleotide long fragments. The data are shown in Figure 2. This analysis of the repetitive fraction provides a more accurate estimate of the repetitive reaction rate constant. About half of the repetitive reaction takes place with a rate constant of 3.0 liter-sec/mole, corresponding to a

Figure 2: Renaturation of selected repetitive sequences

350 nucleotide ^{3}H labeled rat DNA was renatured to Cot 100 and the double strand containing fraction (30%) separated on hydroxyapatite. This fraction was driven by unfractionated 350 nucleotide unlabeled rat DNA. The mass ratio of driver to tracer DNA was 100 from Cot 0.001 to Cot 0.5 and 1000 from Cot 1.0 to Cot 1000.

The best least squares fit to the data gives 0.45 of the DNA reannealing with a rate of 2.97 liter-sec/mole and 0.23 reannealing with a rate of 0.00412. 0.12 of the DNA had not reacted at the highest Cots. These rate values were used for the third fit in Table I. The line drawn through the data displays this fit. The two repetitive components are also shown. The line above the data is the whole rat DNA fit from Figure 1.


FRACTION DNA BOUND TO HYDROXYAPATITE
0.0
0.2
0.4
0.6
0.8
1.0
0.001
0.01
0.1
1.0
10
100
1000
10,000
100,000
EQUIVALENT C_0t

repetition frequency of 10,000. The remainder hybridizes with a rate constant of 0.004 liter-sec/mole, corresponding to a repetition frequency of 10. These values can be used in the least squares fit of the whole rat data and provide a third interpretation of the kinetic fractions in the rat genome in Table I. There is little difference in the goodness of fit criterion for each of the three fits, and we will use the second interpretation (B) as our best model for rat DNA renaturation.

Ten percent of rat DNA has renatured by the earliest times shown on the curve. These sequences may contain very highly repeated (>500,000 fold) DNA sequences and fold-back self-complementary sequences. Again, some of this fraction is due to single strand tails on duplexes bound to HAP.

This paper will concentrate on the repeated fraction of the genome renaturing before Cot 50. It is difficult to study the low (10 to 50 fold) repeat fraction because of substantial contamination by single copy sequences. These sequences have been classed with non-repeated fractions for the purposes of this paper. The rapidly reannealing sequences are included in most of our analysis. There is evidence (Wu, et. al. in prep.) that the fold-back complementary repeat sequences are moderately repeated sequences similar to the other repeated sequences in the genome.

Melting experiments

The number and length of the repeated sequences can be calculated from measurements on the fraction of the genome in true duplex after renaturation to repetitive Cots. We have used optical melting and S-1 nuclease digestion to measure the fraction of the genome containing repeated sequences and the length of the duplexed repeats.

To measure the repeated fraction of rat DNA optically, 1000 nucleotide long fragments were renatured to Cot 5 and Cot 50 and melted in a spectrophometer. Figure 3 shows a sample melt, which includes a native DNA standard and a remelt of the melted DNA. The remelt shows the contribution of single strand tails and very rapidly reannealing sequences to the hyperchromicity of the duplex. Table II summarizes the measurements on Cot 5 and Cot 50 duplexes. These values are very sensitive to corrections for single strand hyperchromicity. In addition, some of the apparent single strand hyperchromicity is due to renaturation of rapidly reannealing sequences indicated by the transition above 80 ^{o}C. The hyperchromicity in the transition, 0.015 to 0.02 is consistent with 5% rapidly reannealing DNA. When this 5% fraction is added to the values in Table II, we find 18% of the genome is in duplex at Cot 5 and 24% is in duplex at Cot 50.

An average length for repeat duplexes can be measured by melting DNA fragments of different length containing duplexes (Graham et. al., 1974). DNAs of different lengths were renatured to Cot 50 and duplex containing strands isolated on hydroxyapatite. The optical melt of the duplex containing fragments is shown in Figure 4 and the results are summarized in Table III. The hyperchromicity of each strand length is a

Figure 3: Melt of rat DNA fragments renatured to Cot 5.

DNA was sheared to 1000 nucleotides, denatured and renatured to Cot 5 in .12 M PB. The sample was then melted in a spectrophotometer equiped with a thermal cuvette. The temperature was raised at a rate of 0.5 ^{o}C/minute to 98 ^{o}C. After melting, the samples were cooled to 60 ^{o}C and remelted.

(□) 1000 nucleotide DNA renatured to Cot 5. (+) remelt of the DNA. (◇) native DNA.

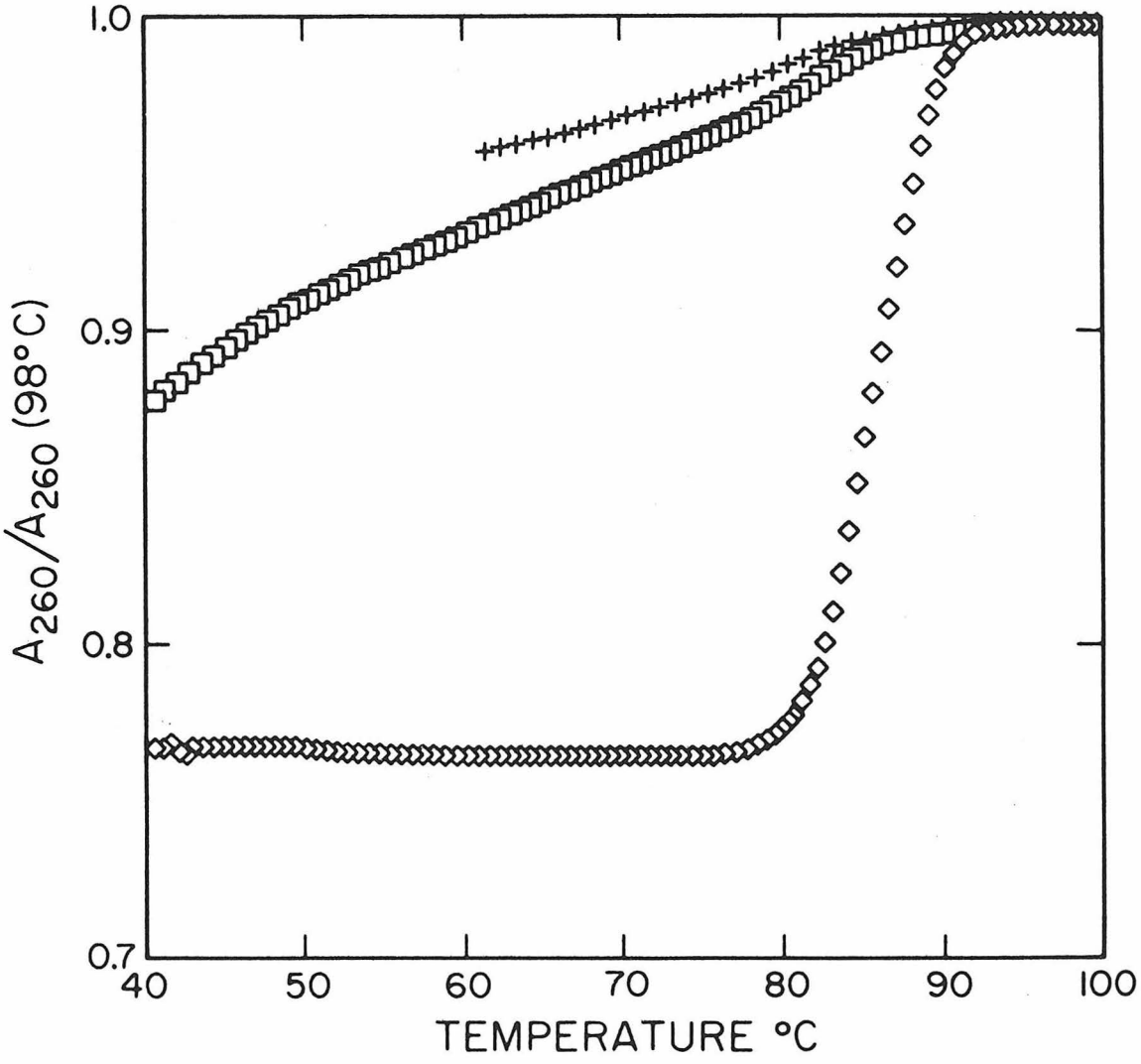

1.0
0.9
0.8
0.7
A_{260}/A_{260} (98°C)
40
50
60
70
80
90
100
TEMPERATURE °C

Table II

Melts of 1000 nucleotide rat DNA fragments

Renatured to Cot 5 and Cot 50

Sample	Hyper-chromicity[a]	T_m °C	Fraction[b] native
Cot 5A	0.0800	77.	0.1818
B	0.0687	75.5	0.1305
C	0.0586	75.	0.0764
D	0.0677	76.	0.1259
Average fraction native 0.1287 ± 0.0431			
Cot 50A	0.0788	74.5	0.1764
B	0.0821	75.5	0.1914

Footnotes to Table II

[a] Hyperchromicity measured from 60 °C to 98 °C.

[b] Fraction native is corrected for 0.04 hyperchromicity after remelt and uses 0.2600 for the native hyperchromicity. The calculation is:

$$\frac{\text{Hyper.} - 0.04}{0.2600 - 0.04}$$

Figure 4: Melt of rat DNA fragments containing a repeated sequence.

DNA was sheared to three lengths, 350, 1100 and 1600 nucleotides determined by alkaline sucrose. The samples were then renatured to Cot 50 and passed over hydroxyapatite. A double strand fraction eluted by 0.5 M PB was dialysed to 0.12 M PB and melted as in Figure 3. A melt of native DNA is included in the figure as a reference.

(+) 350 nucleotide DNA fragments. (□) 1100 nucleotide DNA fragments. (×) 1600 DNA nucleotide fragments. (◇) native DNA.

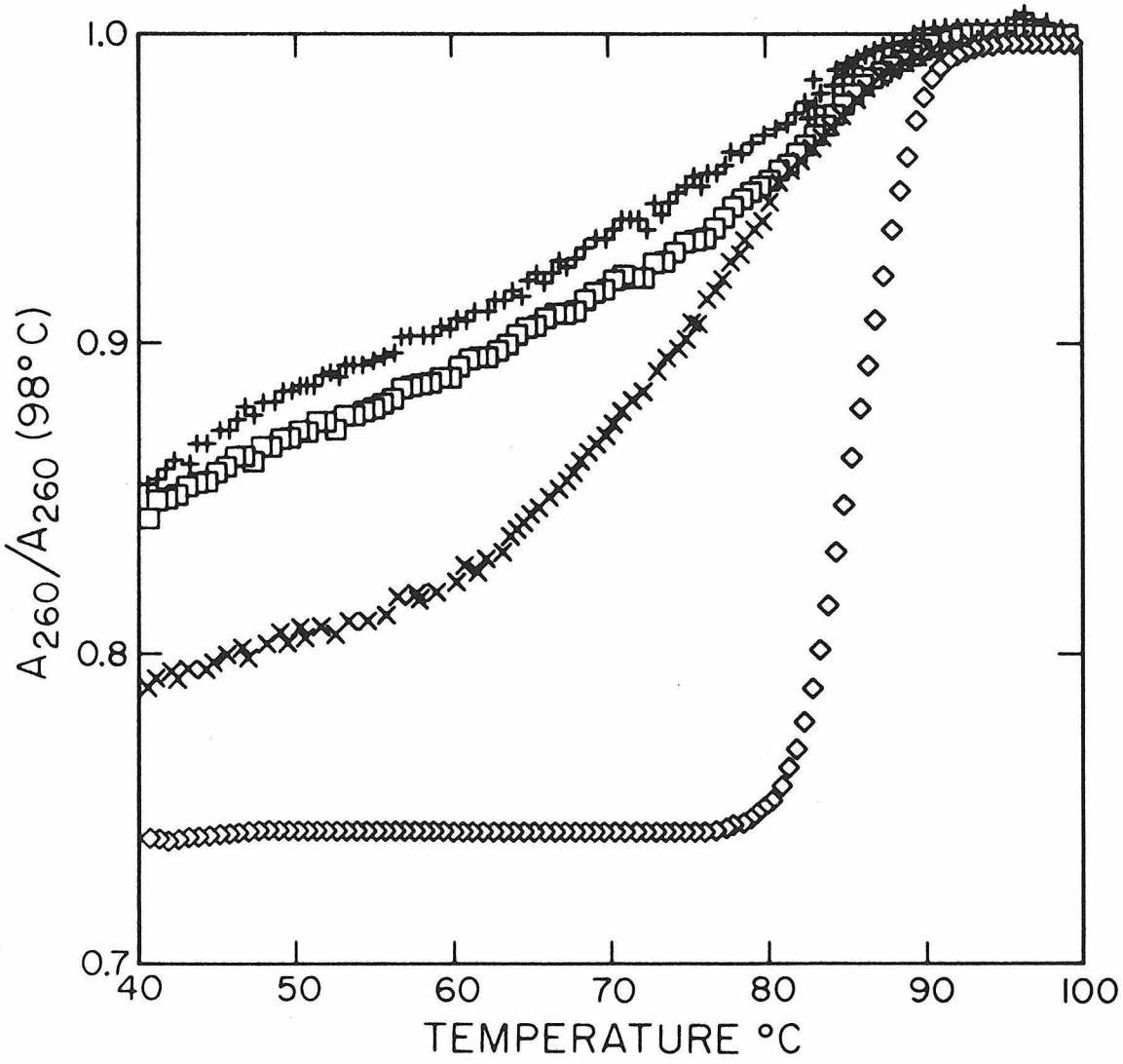

1.0
0.9
0.8
0.7
A_{260}/A_{260} (98°C)
40
50
60
70
80
90
100
TEMPERATURE °C

Table III

Melts of rat DNA fragments

Bearing repeated sequences

Input fragment length	350	1100	1600	Native
Fraction bound to HAP	0.26	0.44	0.47	
Hyper-chromicity	0.175	0.110	0.090	0.260
Fraction duplex[a]	0.646	0.375	0.292	
Duplex length[b]	226	413	467	
Fraction of genome in duplex[c]	0.168	0.165	0.137	

Footnotes to Table III

[a] Fraction duplex is calculated as in Table II for fraction native.

[b] Duplex length is the product of the fraction duplex and the input fragment length

[c] Fraction genome in duplex is the product of the fraction duplex and the fraction bound to HAP (hydroxyapatite)

function of the average length of the strand in duplex. Longer strands have lower hyperchromicity because the duplex region is a smaller fraction of the molecule (Davidson et. al., 1973b). These measurements indicate the average length of the duplex is 400 nucleotides.

These data also provide an estimate of the amount of repetitive sequences in the genome. Although not all of the duplex molecules can be removed from hydroxyapatite by elution with the 0.5 M phosphate buffer used in these experiments, if the fraction retained has the same hyperchromicity as the duplex fraction eluted, about 17% of the DNA is duplex at Cot 50.

Nuclease digestion of repeated sequences

An independent measure of the fraction duplex is provided by digestion with the single strand specific nuclease S-1. Fragments 1000 nucleotides long were renatured to Cots from 0.05 to 50 and duplexed molecules fractionated on hydroxyapatite with and without S-1 nuclease digestion (Table IV). The difference between binding with and without nuclease is due to the absence or presence of single strand tails. Using the S-1 criterion, 6% of the DNA contains repeats renaturing by Cot 0.05 while 16% is repetitive at Cot 5 and 24% at Cot 50.

To make certain that the enzyme was behaving properly, the fraction of duplex as a function of enzyme to DNA ratio was investigated at Cot 5. The 16% of the DNA in duplex at Cot 5 is not a function of enzyme concentration over the range 15 ul/mg to 30 ul/mg (15 ul/mg is the standard concentration used for a DIG = .78; Britten et. al., 1977) and is not affected by a two fold increase in the salt concentration.

Table IV

Nuclease digestion of 1000 nucleotide DNA fragments

Cot	Fraction containing duplex before digestion[a]	Fraction duplex after S-1 digestion[b]
0.05	0.215 ±0.017 (n=2)	0.080 ±0.046 (n=3)
0.5	0.307 ±0.028 (n=2)	0.134 ±0.032 (n=2)
5.	0.282 ±0.032 (n=3)	0.156 ±0.031 (n=5)
50.	0.403 ±0.051 (n=5)	0.239 ±0.010 (n=4)

Footnotes to Table IV

[a] Fraction of DNA bound to hydroxyapatite before S-1 nuclease digestion

[b] Fraction of DNA bound to hydroxyapatite after 45 min digestion with S-1 nuclease (15ul/mg, DIG=0.78 Britten et. al., 1977)

Figure 5: Profile of rat repeated DNA duplexes on agarose A-50

DNA sheared to 3000 nucleotides was denatured and renatured to Cot 5, digested with S-1 nuclease and bound to hydroxyapatie. The double strand fraction (15%) was eluted with 0.5 M PB and chromatographed on Bio-gel agarose A-50.

The size of the fraction indicated was determined by alkaline sucrose sedimentation. The fractions marked were used for the melt in Figure 6.

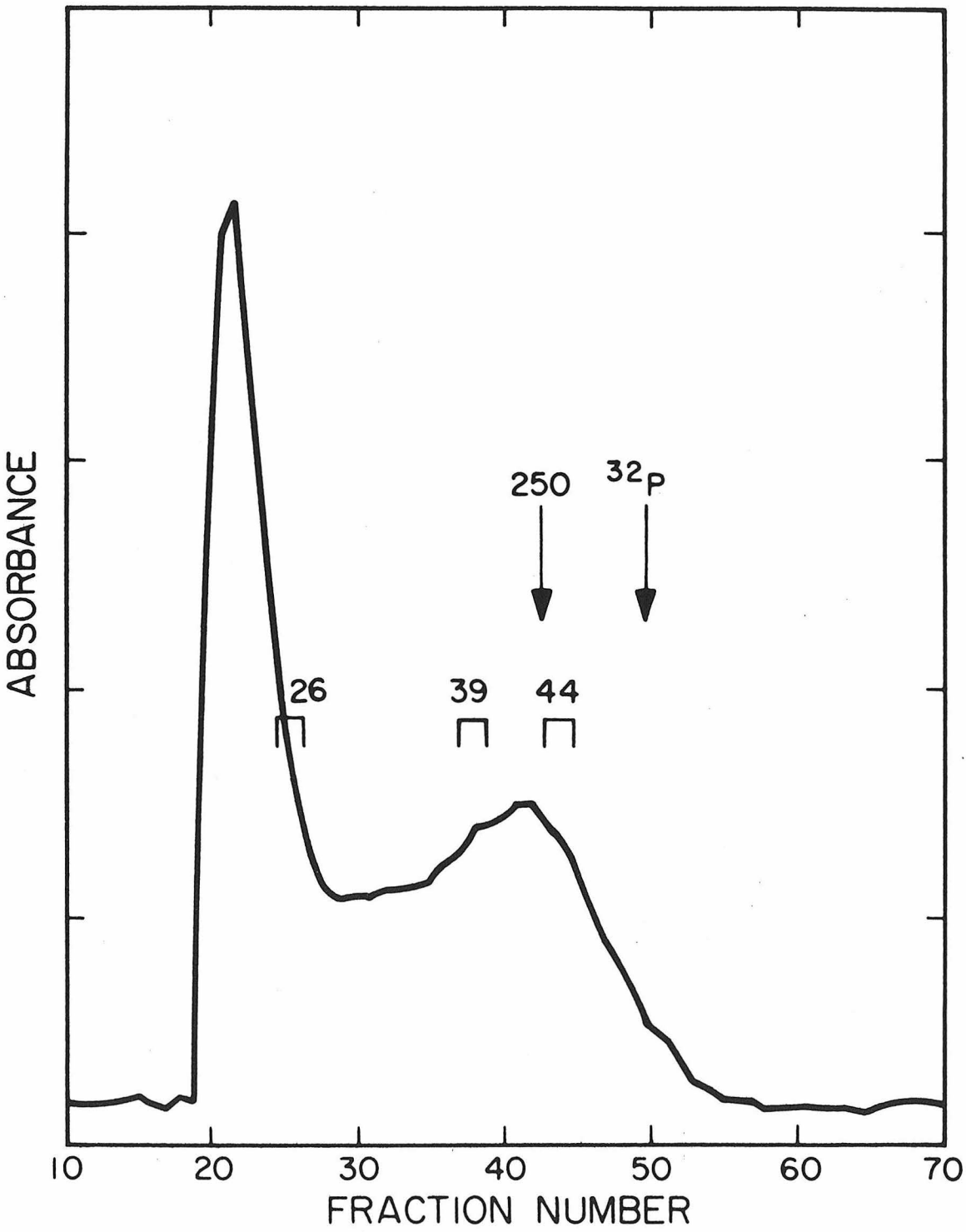

ABSORBANCE
250
32P
26
39
44
10
20
30
40
50
60
70
FRACTION NUMBER

Table V

A-50 fractionation of repeated DNA fragments

Cot	Fraction S-1 resistant	Fraction excluded from A-50
0.05	0.0368	0.484
	0.0558	0.683
5.0	0.150	0.453
	0.140	0.489
	0.171	0.481
	0.167	0.553
	0.177	0.534

Lowering the enzyme concentration or raising the salt concentration should stabilize poorly matched duplexes. We do not believe the enzyme is digesting slightly mismatched regions as these changes do not affect our results.

Fragments from nuclease digestion can be used to determine the length of the duplex by direct sizing. Duplexes from 3000 nucleotide DNA fragments isolated on hydroxyapatite after nuclease digestion were fractionated by size on Biogel agarose A-50. Figure 5 shows a typical elution pattern for this DNA renatured to Cot 5. The included and excluded material was sized on alkaline sucrose gradients and by electron microscopy. The material in the included peak is 200-300 nucleotides long while the excluded material is longer than 1500 nucleotides. Table V summarizes hydroxyapatite and A-50 fractionation data at Cot 0.05 and Cot 5.

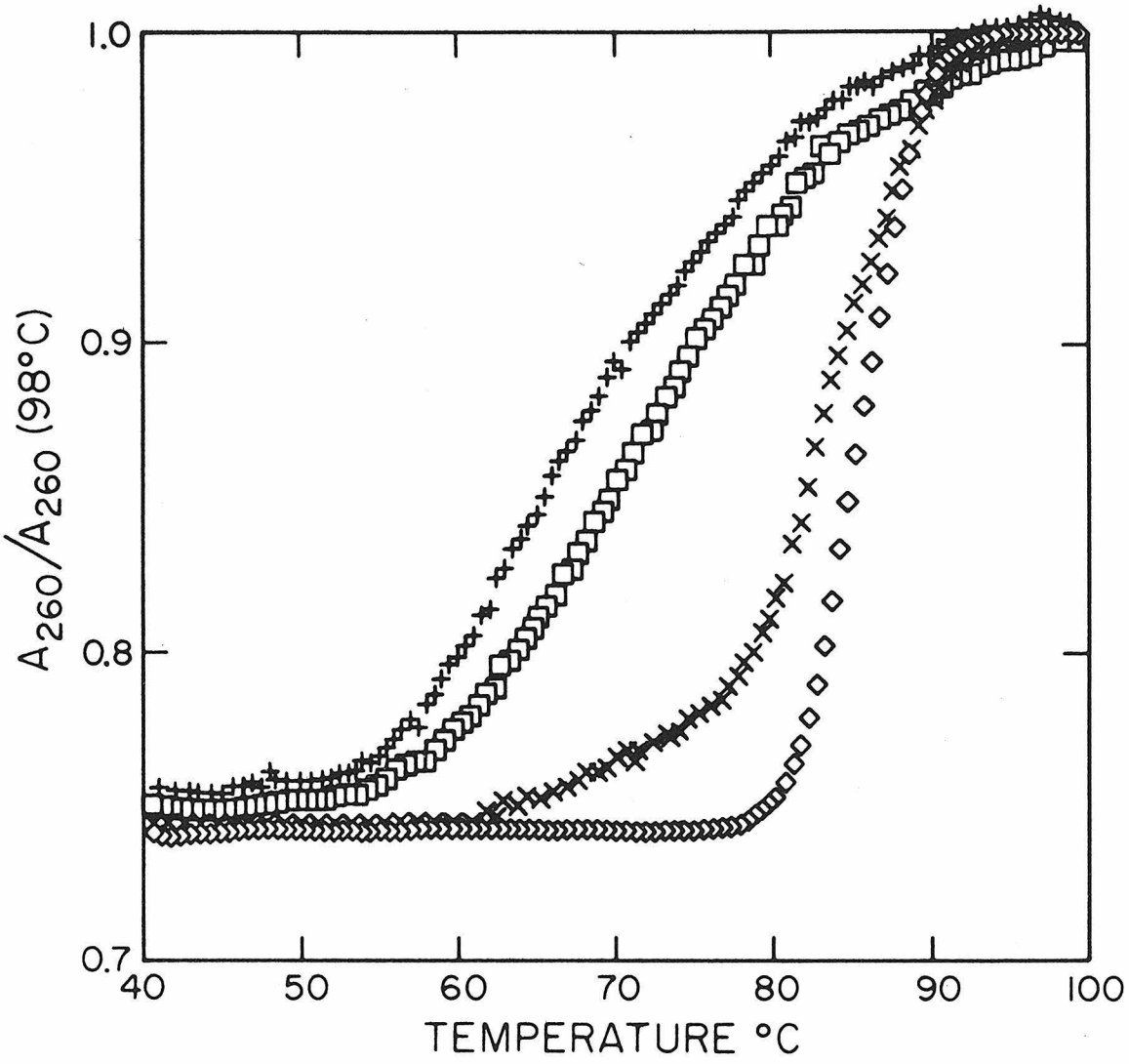
To make certain that long duplexes did not contain single strand tails, duplexes fractionated on A-50 were melted. A sample melt of three fractions indicated in Figure 5 is shown in Figure 6. All fractions show more than 90% of native hyperchromicity. The melting temperatures range from 2.5 ^{o}C below native for the excluded material to 15^{o}C below for the 200 nucleotide fragments.

The melts of both the long and short repetitive fragments display a number of distinct melting regions. These regions can be seen most clearly in the melt of the long A-50 excluded material where there are components melting with T_m's of 70 ^{o}C, 82 ^{o}C and 86 ^{o}C. The short fragments also show components with T_m's of 65 ^{o}C, 75 ^{o}C and 83 ^{o}C.

Figure 6: Melt of repeated DNA duplexes of different fragment lengths.

Repeated sequence DNA duplexes were isolated and fractionated as described in Figure 5. Fractions 26, 39 and 44 were melted as described in Figure 3.

(+) fraction 44; (□) fraction 39; (×) fraction 26; (◇) native DNA reference.


1.0
0.9
0.8
0.7
A_{260}/A_{260} (98°C)
40
50
60
70
80
90
100
TEMPERATURE °C

Interspersion of repeated sequences

The data on melting and nuclease digestion of long DNA fragments presented earlier indicate that many short repeated sequences are surrounded by single copy DNA. In this section we present two measures of the quantity of interspersed repeated sequences. The first experiment gives a qualitative measure of the fraction of the genome containing interspersed repeats; the second gives quantitative data on the fraction of the DNA interspersed, the length of the interspersed single copy sequences and the length of the repeated sequences.

When 300, 1500 and 3000 nucleotide long DNA fragments are renatured and the duplex containing fraction separated on hydroxyapatite, the repeated fraction appears to increase with the length of the renatured fragments. Figure 7 shows the renaturation of these three fragment lengths. Thirty percent of 300 nucleotide long fragments are bound at Cot 50 but 70% of 3000 nucleotide DNA fragments are retained on hydroxyapatite at this Cot. The increased binding is due to single strand tails on short duplexes. These tails were seen as loss of hyperchromicity in the melting experiments and loss of hydroxyapatite binding after S-1 nuclease digestion.

The precise fraction of the DNA containing short interspersed repeats and their spacing can be measured by hybridizing trace quantities of varying length DNA fragments with a vast excess of unlabeled short DNA. Whole rat DNA sheared to 350 nucleotides was used to drive labeled fragments from 250 to 8000 nucleotides in length. The fraction bound to hydroxyapatite at zero time (<Cot 0.005), Cot 5 and Cot 50 was measured. Figures 8A and 8B show the binding of the DNA at

Figure 7: Renaturation of different DNA fragment lengths

DNA sheared to 300, 1500 and 3000 nucleotides determined by alkaline sucrose sedimentation was denatured and renatured. The samples were passed over hydroxyapatite and the double strand containing fraction eluted at 100 ^{o}C with .12 M PB. The fraction single stranded as a function of equivalent Cot is plotted.

(O) 300 nucleotide DNA; (□) 1500 nucleotide fragments; (△) 3000 nucleotide DNA fragments.

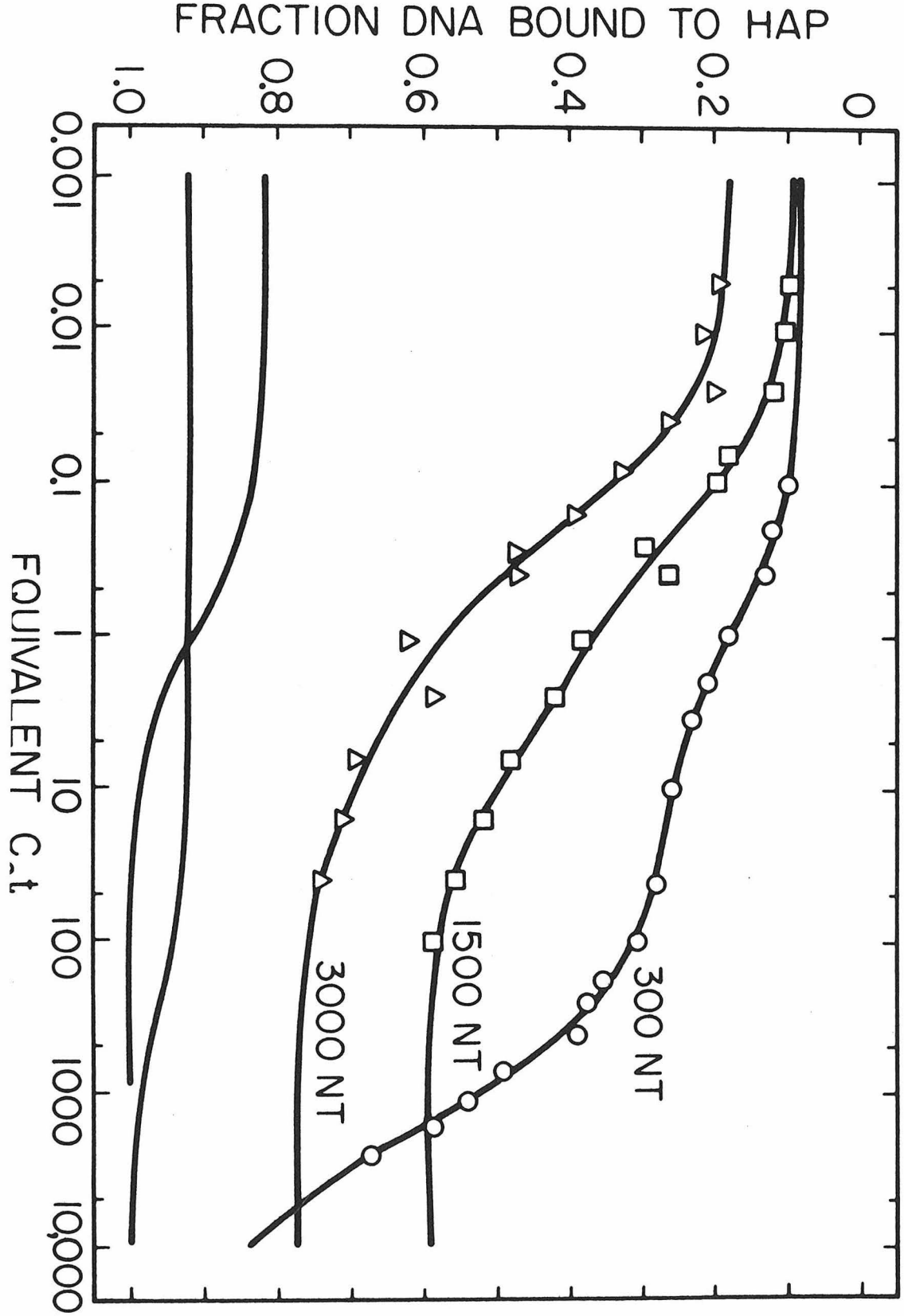

FRACTION DNA BOUND TO HAP
1.0
0.8
0.6
0.4
0.2
0
0.001
0.01
0.1
1
10
100
1000
10,000
FQUIVALENT C.t
3000 NT
1500 NT
300 NT

Figure 8: The fraction of rat DNA sequences containing a repeated DNA sequence as a function of fragment length.

Labeled DNA fragments of various sizes were driven by whole rat DNA to Cot 5 and Cot 50. Binding to hydroxyapatite without driver at very short times (zero time binding) was also measured. The fraction of the fragments containing duplex F was measured by hydroxyapatite chromatography. Values of F were corrected for the fraction of zero time binding sequences Z. The zero time binding correction was calculated from a linear fit of the data in Figure 8C. The values plotted in Figures 8A and B are

$$R = \frac{F - Z}{1.0 - Z}$$

The solid line drawn to the data represents the best least squares fit to the data. The dashed lines provide alternative interpretations of the data. A summary of the interpretations is given in the text.

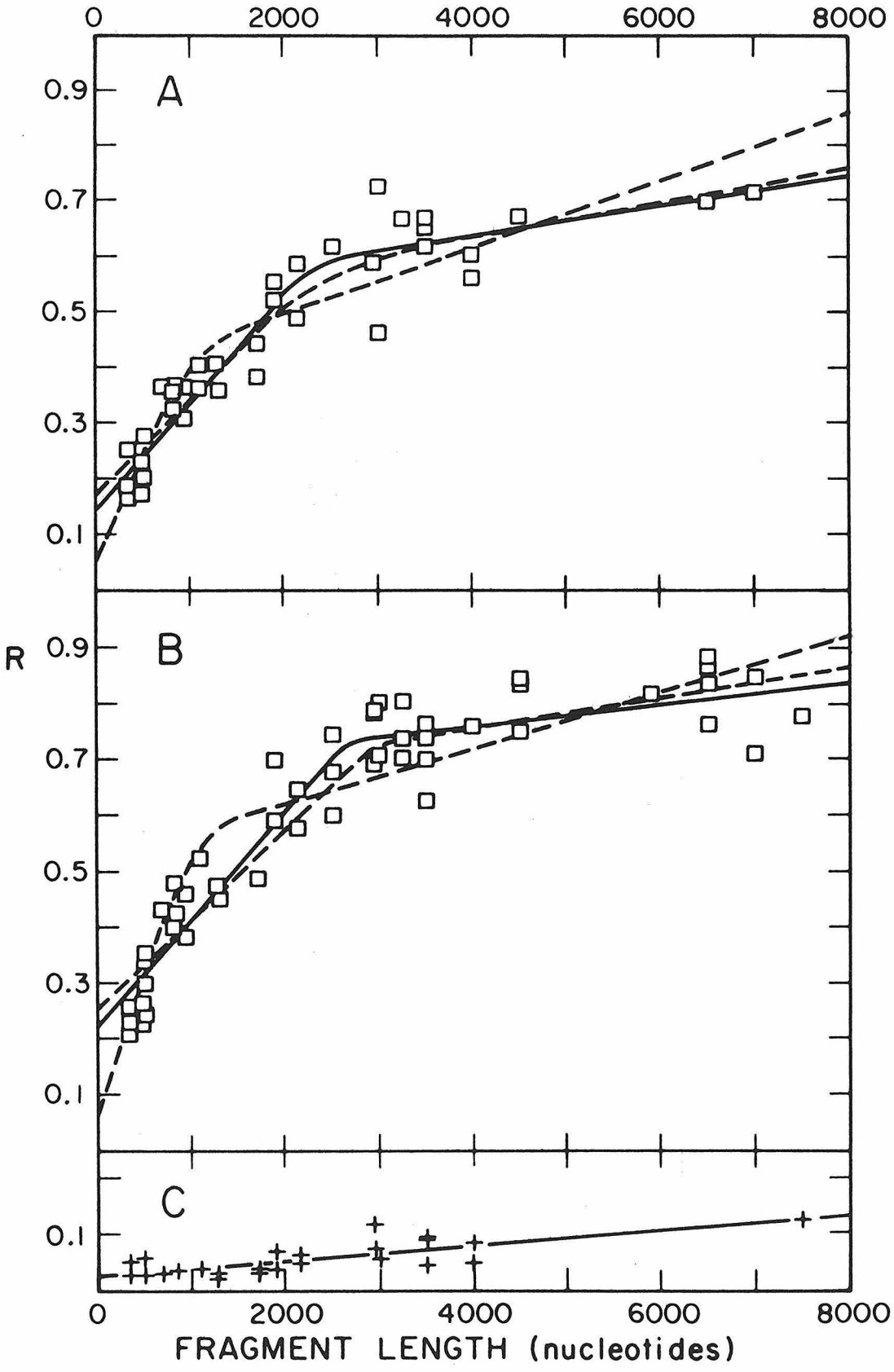
8A: Corrected binding of fragments driven to Cot 5. Solid line: best fit with fraction repetitive (f_{rep}) = 0.140, interspersion period (L_{int}) = 2315 nucleotides (NT), fraction interspersed at short period (f_{int}) = 0.573, repetitive sequence length (L_{rep}) = 835 NT, fraction bound by 8000 nucleotides (f_{8000}) = 0.745.

One dashed line shows the fit when the fraction repetitive is fixed at 0.17 while the other shows the best fit of the data when the interspersion distance is set to 1000 NT.

8B: Corrected binding of fragments driven to Cot 50. Solid line: best fit with f_{rep} = 0.219, L_{int} = 2680, f_{int} = .730, L_{rep} = 1270 NT, f_{8000} = .836.

The dashed lines show the best fits when the fraction repetitive is set to 0.25 and the interspersion period is set to 1000 NT.

8C: Binding of fragments incubated without driver or for very short times (zero time binding). The line drawn to the data was used to correct the data in Figures 8A and 8B.


0
2000
4000
6000
8000
A
B
C
R
0.9
0.7
0.5
0.3
0.1
FRAGMENT LENGTH (nucleotides)

Cot 5 and Cot 50 corrected for the zero time binding fraction. Figure 8C shows the zero time binding used to correct the data. The lines plotted show the range of curves which can be used to fit the data. The curves show that more than 80% of the 8000 nucleotide DNA contains repetitive sequences hybridizing by Cot 50. The spacing between repeats (the interspersion period) is about 2300-2700 nucleotides and does not change significantly between Cot 5 and Cot 50. At Cot 5, 57% of the DNA is bound at a fragment length of 2300 nucleotides; 73% of the DNA is bound at a fragment length of 2700 nucleotides at Cot 50. The fraction of the DNA in repeated duplexes is the Y-intercept of the data, about 14% at Cot 5 and 22% at Cot 50. A length estimate for the repeat duplex is shown at the X-axis intercept and is about 835 nucleotides at Cot 5 and 1300 nucleotides at Cot 50.

Discussion

We have presented measurements on the repetition frequency, fraction of the genome and length of repeated DNA sequences in rat DNA. In this section we will summarize the different lines of evidence which support a consistent model for repeated sequence organization in the rat genome.

The kinetic data from the hydroxyapatite Cot curve are consistent with the reported genome size for the rat (2.9 pg, Sober, 1968). A variety of combinations of rate constants and repetitive and single copy fraction quantities describe the data equally well. This is indicated by the standard error of the parameters in Table I. Some of this

uncertainty is due to the difficulty of getting good termination data for the reaction. A change from a final fraction unreacted from 5% to 1% changes the apparent rate constant for single copy reaction from 0.00032 to 0.00024 liter-sec/mole. In addition, the presence of a ten-fold repeated fraction makes accurate determination of the single copy rate constant more difficult.

The slave "mini-Cot" reaction (Britten et. al., 1974) provides more accurate information about the moderately repeated fraction. Isolation of the fraction excludes many of the sequences repeated less than 10 times, however. The mini-Cot analysis certainly suggests there are two definite frequency classes of repeated DNAs.

Measurements on repeated sequence elements

The number, size and spacing of the repeated sequences have been measured by melting, nuclease digestion and interspersion experiments. Each of these experiments measures one number well, and provides other information which should be consistent. In this section we will try to collate the data from each of the experiments to find agreement on the fraction of the genome which is repetitive and on the length of the repeated sequences.

The fraction repetitive

Melting presents the best data on the fraction repetitive. Nuclease digestion may destroy mismatched duplexes which are considered repeated by standard hydroxyapatite fractionation. The interspersion data do not give precise values for the fraction repetitive or length of

the repeated sequences. Melting data are difficult to interpret because of the large contribution of collapsed single strand tails when only a small fraction (<20%) of the DNA is duplexed. We have tried to control for the problem by immediately remelting the samples and subtracting the hyperchromicity due to single strands. Some of the single strand hyperchromicity is due to rapidly reannealing sequences, however, and the simple calculation:

$$\frac{\text{single strand hyperchromicity - duplex hyper.}}{\text{single strand hyper. - native hyper.}}$$

gives 13% duplex at Cot 5 and 19% at cot 50. The single strand hyperchromicity corrections were made at 60 °C; this excludes contributions from very short duplexes which melt below 50 °C. A 5% rapidly renaturing fraction subtracted out by the single strand correction must be added to the 13% and 19% to find the total fraction in duplex: 18% at Cot 5 and 24% at Cot 50.

These numbers agree closely with the nuclease values presented in Table IV. The enzyme ratios used in these experiments selectively clip single strand tails and do not attack slightly mismatched duplexes (Britten et. al., 1977). Our studies with a range of enzyme ratios and salt concentrations suggest that poorly matched duplexes are exposed to the same criterion with nuclease digestion as with optical melts and hydroxyapatite fractionation. The interspersion curves at Cot 5 and Cot 50 are also consistent with 18% of the genome in duplex at Cot 5 and 24% in duplex at Cot 50.

Repetitive sequence length

The best measure of repetitive sequence length comes from nuclease digestion studies. These studies display the real distribution of sizes on agarose A-50. About 40% to 50% of the repeated DNA sequences are excluded from A-50 while the remainder are included with a length of 200-300 base pairs. These lengths are similar to results from the sea urchin (Davidson et. al., 1973b; Britten et. al., 1976), Aplysia (Angerer et. al., 1975), a host of marine invertebrates (Goldberg et. al., 1975) and an insect (Efstratiadis et. al., 1976). The distribution of excluded/included material is variable. This distribution may be much more sensitive to enzyme concentration than the fraction of DNA bound to hydroxyapatite and in some cases excluded material may include short sequences attached by poorly matched, nuclease susceptible regions. The variability may also be due to complementary long and short sequences. Short duplex hybrids on long molecules may have a much higher sensitivity to slight changes in incubation conditions.

If 60% of the mass of the repeated sequences are 275 nucleotides long and the remainder 1500 nucleotides long, the weight average length is (0.6*275)+(0.4*1500) = 765.0 nucleotides. Under ideal conditions the duplex melting experiment would provide the weight average number while the interspersion data would provide a number average estimate. The number from the melting experiments ranges from 225 nucleotides duplex length on 350 nucleotide fragments to 470 nucleotides duplex on 1600 nucleotide fragments. This may be due to the relatively short (300, 1100 and 1600 nucleotide) DNA fragment lengths used in the experiment.

On the average, even the longest fragment length, 1600 nucleotides, could only form an 800 nucleotide duplex (Smith et. al., 1975) when two long repetitive sequences renature. Long repetitive sequences would not account for their full fraction of hyperchromicity until fragments longer than 3000 nucleotides were reacted. Such reactions are technically difficult; 3000 nucleotide long fragments form duplexes which cannot be eluted from hydroxyapatite with 0.5 M phosphate and must be denatured.

The interspersion long tracer/short driver experiments provide a number average repetitive sequence length near 1000 nucleotides. It is difficult to interpret these values; the extrapolation of the interspersion data to zero fraction bound is dependent on assumptions about the interspersion of the long fragments.

A consistent model for repeated sequences must explain the 350 nucleotide hydroxyapatite binding data. According to these data, 24% of the fragments contain duplex at Cot 5 and 31% contain duplexes at Cot 50. This contrasts with 18% duplex at Cot 5 and 24% duplex at Cot 50 measured by melting and nuclease digestion. This implies 75% of the strands bound on hydroxyapatite are duplexed, in agreement with the melting data of Figure 4 (65% in duplex). If 0.6 of the repeated sequences are about the same length as the renatured fragments, (300 nucleotides from A-50 fractionation) 50% of those fragments should be duplexed due to random overlap (Smith et. al., 1975; Britten and Davidson, 1976). The same argument gives 1350/1500 or 90% of the 1500 long fragments in duplex. This implies 0.6*0.5 + 0.4*0.9 = 0.66 of the strands should be duplexed. The agreement between the 75% fraction

duplex after hydroxyapatite binding and the 66% expected from the repetitive sequence length is good. The more repeated sequences will be more than 50% covered but we may have overestimated the fraction of short repeats. In any case the hydroxyapatite data are in excellent agreement with our estimates for the fraction and length of the repetitive duplex.

These estimates for sequence length are also similar to measurements made on 1000-3000 nucleotide fragments renatured and visualized in the electron microscope (Bonner et. al., 1973; Wilkes et. al., 1977). Three lengths of DNA were stripped of rapidly renaturing sequences and renatured to Cot 50. The duplexes formed from 900 nucleotide fragments had a number average length of 329 nucleotides and a weight average of 542 nucleotides. Duplexes formed from 1700 nucleotide strands had number and weight average lengths of 376 and 675 nucleotides, while 450 and 990 nucleotide duplexes were formed by 2500 nucleotide strands. The increase in average duplex length with fragment length is due to the long repeated sequences. Longer fragments can form longer structures identifiable as duplexes in the electron microscope. Duplexes formed by 2500 nucleotide fragments have a weight average close to the 770 nucleotides predicted from our nuclease experiments. Duplexes formed by shorter fragments display the length of the short repeat fraction, 200-400 nucleotides. Electron microscopy experiments by Chamberlin et. al. (1975) find the short repetitive sequence length in Xenopus is 350 nucleotides, in good agreement with our findings. The fragment size used the Xenopus experiments precluded measurements on long duplex sequences.

While the data on the fraction of long and short repetitives (Table V) are variable, we believe they reflect distinct lengths of repetitive sequences in rat DNA. Our numbers suggest the rat genome can be separated into the repetitive and single copy classes shown in Table VI. Seventy percent of the DNA is single copy, 25% repeated from 100-100,000 fold and 5% contains sequences in duplex before Cot 0.05. The 16% in duplex by Cot 5 can be divided into a long fraction with 6.4% of rat DNA and a short 200-300 nucleotide fraction with 9.6% of the DNA.

The long sequences differ from shorts not only in size but also in fidelity of matching. The melting data in Figure 6 show the same melting temperature dependence on length reported by Davidson et. al. (1973b), and Goldberg et. al. (1975). The T_m differences between long and short repeated sequences are not simply due to the differences in length. The 650/n equation expressing dT_m as a function of nucleotide length n (Britten et. al., 1974) would only predict a 3 degree decrease in T_m for 200 nucleotide sequences while the differences are 15 degrees. This corresponds to 12% mismatch in short repeats assuming 1.0% mismatch per degree decrease in T_m (Britten et. al., 1974). The long and short fragment melting curves show more than one transition however. There are long sequences with low precision regions and high precision short repeated sequences.

Interspersion of repeated sequences in rat DNA

Almost every one of the results we have presented shows qualitative evidence for interspersion of short repeated sequences with single copy DNA in the rat. If repeat sequences were not interspersed, we could not

measure the duplex length by digesting away adjacent single strands. This is not a trivial result; similar experiments by Crain et. al. (1976) including S-1 nuclease digestion and melting failed to detect interspersed repetitive sequences in Drosophila.

The best data on repeated sequence interspersion come from the short driver/long labeled tracer experiments shown in Figure 8. The range of least squares fits which can be imposed on the data is indicated in the figure. Changes in the fraction repetitive move the inflection point for the short interspersion period. When the nuclease and melting data are used to constrain the fit to 16% and 25% repetitive at zero single strand tail length, the short interspersion periods are 2400 and 2700 nucleotides at Cot 5 and Cot 50. The similarity of interspersion periods at those two Cots suggests different frequency repeats are probably not arranged differently. At these interspersion periods, 57% and 73% of the DNA is organized as short interspersed repeats while another 15% and 10% of the DNA is interspersed with a period longer than 6000 nucleotides.

Electron microscopy experiments mapping short duplex fragments on very long (10,000-50,000 nucleotides) DNA have also been used to determine the interspersion period in the rat (Bonner et al., 1973, Wilkes et. al., 1977). These data cannot show the same division of interspersion periods shown in Figure 8 but show a large class of spacings between 1000 and 3000 nucleotides.

These data are in excellent agreement with measurements on human DNA (Schmid and Deininger, 1976) and are not significantly different from measurements on Xenopus (Davidson et. al., 1973a), sea urchin

(Graham et. al., 1974) or Aplysia (Angerer et. al., 1975). The data in Figure 8 are open to a range of interpretations. A model of interspersion in the rat which included three interspersion periods, at 800, 4000 and >6000, nucleotides would fit the data equally well.

Conclusions

Table VI presents a model for the number and length of the different sequence fractions in rat DNA. About 70% of the DNA is single copy, the exact fraction is obscured by the low frequency (10-fold) repeated fraction of the DNA seen in Figures 1 and 2. We estimate the slow repeat fraction contains 10% of the DNA. The major repetitive component is repeated about 3000 fold and includes about 20% of the DNA. From S-1 nuclease digestion and agarose chromatography, we can measure the length of the repeated sequences. Sixty percent of the moderately repetitive sequences are 300 nucleotides long and 40% are longer than 1500 nucleotides. From this length distribution we calculate there are 50 different families of sequences 1500 nucleotides long repeated 3000-fold and 350 different 300 nucleotide sequences with the same repetition frequency.

The interspersion measurements also allow us to calculate the length and number of the single copy sequences. There are 560,000 single copy sequences 2700 nucleotides long and another 50,000 sequences longer than 6000 nucleotides. There are enough short repeated sequences (1,050,000) to bound each of the short single copy sequences.

Table VI

Sequence distribution in rat DNA

Comp	Fraction of DNA[a]	Size Distribution	Repetition frequency	Complexity (nucleotides)	Number of Sequences
1	0.02[b]		>100,000	530.	
2	0.20[c]	0.4 > 1500 0.6 200-400	3000	1.76×10^5	50 350
3	0.08[d]		12.5	1.69×10^7	
4	0.70	0.75 2500 0.25 >6000	1.	1.85×10^9	560,000 <80,000

Footnotes to Table VI

[a] Fraction of DNA is our best estimate of the true fraction duplex. These values differ from those in Table I because of single strand tails bound to hydroxyapatite.

[b] This fraction assumes that .67 of the 0.06 of the duplex bound before Cot 0.05 is 3000 fold repeated sequence (Wu et. al., in prep.)

[c] This fraction includes 0.04 from the binding before Cot 0.05.

[d] This is a rough estimate. None of the nuclease digestion studies separated the low repeat fraction from single copy DNA. From the hydroxyapatite data of Table I, 0.12 of the genome consists of slow repeats. We have assumed 0.67 of that DNA is true duplex.

The highly structured interspersion of the short repeated DNA sequences, the universality of this pattern of organization and the proximity of expressed sequences to subsets of repeated sequences all suggest an important functional role for repeated sequences. Britten and Davidson (1969) have presented a model for regulatory function of repeated DNA sequences which is consistent with a wide range of structural and functional features of DNA sequences discovered since the model was proposed.

The distribution and function of the long repeated DNA sequences is unclear. Some of this DNA must contain the known repetitive gene families such as the ribosomal and histone genes (Brown and Sugimoto, 1974; Birnsteil et. al., 1974; for discussion see Galau et. al., 1976) and must be distinct from short repeated sequences. Long repeated DNA may contain other sequences which are shared with short repeated sequences. Some repeated sequences may sometimes be bounded by single copy sequences and other times be bounded by other repeated sequences as part of a long repeated sequence. Although we have begun to explore the sequence relationships between long and short repeated DNA sequences (Pearson et. al., 1977b), measurement of sequences shared by these two size classes is technically difficult. Many short fragments may be derived from long repeated sequences because of mechanical shear. Purification of long sequences is easier but it is difficult to design experiments which are insensitive to 5- to 10% contamination by short repeated sequences. If sequence overlap can unambiguously be demonstrated, it may provide evidence for the "integrator" gene sequences postulated by Britten and Davidson (1969).

References cited

Angerer, R. C., Davidson, E. H. and Britten, R. J. (1975) "DNA sequence organization in the mollusc Aplysia californica" Cell 6:29-39

Bonner, J., Garrard, W. T., Gottesfeld, J. M., Holmes, D. S., Sevall, J. S. and Wilkes, M. M. (1973) "Functional organization of the mammalian genome" Cold Spring Harb. Symp. Quant. Biol. 38:303-310

Birnstiel, M., Telford, J., Weinberg, E. and Stafford, D. (1974) "Isolation and some properties of the genes coding for the histone proteins" Proc. Nat. Acad. Sci. US 71:2900-2904

Britten, R. J. and Davidson, E. H. (1969) "Gene regulation in higher cells: A theory" Science 165:349-357

Britten, R. J. and Davidson, E. H. (1976) "Studies on nucleic acid reassociation kinetics: empirical equations describing DNA reassociation" Proc. Nat. Acad. Sci. US 73:415-419

Britten, R. J., Graham, D. E. and Neufeld, B. R. (1974) "Analysis of repeating DNA sequences by reassociation" Methods in Enzymology (L. Grossman and K. Moldave, eds.) Vol 29 part E 363-418

Britten, R. J., Graham, D. E., Eden, F. C., Painchaud, D. M. and Davidson, E. H. (1977) "Evolutionary divergence and length of repetitive sequences in sea urchin DNA" J. Mol. Evol. in press

Brown, D. D. and Sugimoto, K. (1974) "The structure and evolution of ribosomal and 5S DNAs in Xenopus laevis and Xenopus mulleri: the evolution of a gene family" Cold Spring Harb. Symp. Quant. Biol. 38:501-505

Chamberlin, M. E., Britten, R. J. and Davidson, E. H. (1975) "Sequence organization in Xenopus studied by electron microscopy" J. Mol. Biol. 96:317-333

Crain, W. R., Eden, F. C., Pearson, W. R., Davidson, E. H. and Britten, R. J. (1976) "Absence of short period interspersion of repetitive and non-repetitive sequences in the DNA of Drosophila melanogaster" Chromosoma (Berl.) 56:309-326

Dahmus, M. E. and McConnell, D. J. (1969) "Chromosomal ribonucleic acid of rat ascites cells" Biochemistry 8:1524-1534

Davidson, E. H. and Britten, R. J. (1973) "Organization, transcription and regulation in the animal genome" Quart. Rev. Biol. 48:565-613

Davidson, E. H., Hough, B. R., Amenson, C. S. and Britten, R. J. (1973a) "General interspersion of repetitive with non-repetitive sequence elements in the DNA of Xenopus" J. Mol. Biol. 77:1-23

Davidson, E. H., Graham, D. E., Neufeld, B. R., Chamberlin, M. E., Amenson, C. S., Hough, B. R. and Britten, R. J. (1973b) "Arrangement and characterization of repetitive sequence elements in animal DNAs" Cold Spring Harb. Symp. Quant. Biol. 38:295-301

Davidson, E. H., Galau, G. A., Angerer, R. C. and Britten, R. J. (1975a) "Comparative aspects of DNA sequence organization in metazoa" Chromosoma (Berl.) 51:253-259

Davidson, E. H., Hough, B. R., Klein, W. R. and Britten, R. J. (1975b) "Structural genes adjacent to interspersed repetitive DNA sequences" Cell 4:217-238

Efstratiadis, A., Crain, W. R., Britten, R. J., Davidson, E. H. and Kafatos, F. (1976) "DNA sequence organization in the lepidopteran Antherea pernyi" Proc. Nat. Acad. Sci. US 73:2289-2293

Galau, G. A., Chamberlin, M. E., Hough, B. R., Britten, R. J. and Davidson, E. H. (1976) "Evolution of repetitive and non-repetitive DNA" Chap. 12 of Molecular Evolution (F. J. Ayala, ed.) Sunderland, MA: Sinauer Assoc. pp 200-224

Goldberg, R. B., Crain, W. R., Ruderman, J. V., More, G. P., Barnett, J. R., Higgins, R. C., Gelfand, R. A., Galau, G. A., Britten, R. J. and Davidson, E. H. (1975) "DNA sequence organization in the genomes of five marine invertebrates" Chromosoma (Berl.) 51:225-251

Gottesfeld, J. M., Bagi, G., Berg, B. and Bonner, J. (1976) "Sequence composition of the template-active fraction of rat liver chromatin" Biochemistry 15:2472-2483

Graham, D. E., Neufeld, B. R., Davidson, E. H. and Britten, R. J. (1974) "Interspersion of repetitive and non-repetitive DNA sequences in the sea urchin genome" Cell 1:127-137

Holmes, D. S. and Bonner, J. (1974a) "Sequence composition of rat nuclear deoxyribonucleic acid and high molecular weight ribonucleic acid" Biochemistry 13:841-848

Holmes, D. S. and Bonner, J. (1974b) "Interspersion of repetitive and single copy sequences in nuclear ribonucleic acid of high molecular weight" Proc. Nat. Acad. Sci. US 71:1108-1112

Manning, J. E., Schmid, C. W. and Davidson, N. (1975) "Interspersion of repetitive and non-repetitive DNA in the Drosophila melanogaster genome" Cell 4:141-155

Noll, H. (1967) "Characterization of molecules by constant velocity sedimentation" Nature 215:360-363

Pearson, W. R., Davidson, E. H. and Britten, R. J. (1977a) "A program for least squares analysis of reassociation and hybridization data" submitted to Nuc. Acids Res.

Pearson, W. R., Wu., J. R., Posakony, J. and Bonner, J. (1977b) "Analysis of sequences in rat repetitive DNA: A preliminary report" Thesis, California Institute of Technology

Plageman, P. G. W. and Swimm, H. E. (1966) "Replication of menogovirus I: Effect on synthesis of macromolecules by host cell" J. Bact. 91:2317-2326

Plageman, P. G. W. (1969) "RNA synthesis in exponentially growing rat hepatoma cells I. A caution in equating pulse labeled RNA with messenger RNA" Biochim. Biophys. Acta. 182: 46-56

Schmid, C. W. and Deininger, P. L. (1975) "Sequence organization of the human genome" Cell 6:345-358

Smith, M. J., Hough, B. R., Chamberlin, M. E. and Davidson, E. H. (1974) "Repetitive and non-repetitive sequences in sea urchin heterogeneous nuclear RNA" J. Mol. Biol. 85:103-126

Smith, M. J., Britten, R. J. and Davidson, E. H. (1975) "Studies on nucleic acid reassociation kinetics: reactivity of single stranded tails in DNA-DNA renaturation" Proc. Nat. Acad. Sci. US 72:4805-4809

Sober, H. A. (1968) "Deoxyribonucleic acid content per cell of various organisms: Table X Mammals" Handbook of Biochemistry Cleveland: Chemical Rubber Co. p. H-58

Studier, F. W. (1965) "Sedimentation studies on the size and shape of DNA" J. Mol. Biol. 11:373-390

Wilkes, M. M., Pearson, W. R. and Bonner, J. (1977) "Rat DNA sequence analysis by electron microscopy" Thesis, California Institute of Technology

CHAPTER II

ANALYSIS OF SEQUENCES IN RAT REPETITIVE DNA

A PRELIMINARY REPORT

Introduction

Recent studies on repetitive DNA sequence organization (Davidson et. al., 1973a; Goldberg et. al., 1975; Angerer et. al., 1975; Efstratiadis et. al., 1976; Pearson et. al., 1977b.) indicate there are two classes of repeated DNA sequences. In many organisms, 300 nucleotide interspersed sequences comprise 50% to 70% of the repeated DNA while 30% to 50% of the sequences are substantially longer, over 1500 nucleotides. Early studies on the arrangement of repeated DNA sequences (Davidson et. al., 1973a; Chamberlin et. al., 1975; Bonner et. al., 1973) did not suggest long repeated sequences. These long sequences are consistent with a model for regulation of gene expression proposed by Britten and Davidson (1969), however, and may serve the function of "integrator genes" proposed in the model.

The existence of two size classes of repeated sequences poses a number of interesting questions. First, are long and short sequences kinetically different? This question is answered by measuring the repetition frequency and complexity of purified long and short repeated sequences. Another question is perhaps more interesting: Do sequences which appear in short DNA sequences also appear in long sequences? This question is more relevant to the "integrator gene" hypothesis (Britten and Davidson, 1969). If some sequences in long "integrator" sequences are also present thoughout the genome as short interspersed sequences, batteries of control elements can be easily imagined (Britten and Davidson, 1969; Davidson and Britten, 1973).

While these are intriguing questions, substantial technical difficulties hamper studies of large classes of random repetitive sequences. The technical problems stem from the difficulties in isolating pure preparations of long and short repeats. The short repeat fraction must always be contaminated by long sequences digested into short fragments because of overlap errors during renaturation. While one can be confident of much higher purity in the long repeat sequences, it is difficult to design experiments which can exclude the effects of 1% to 10% contamination of short sequences with high repetition frequency. We will examine these problems in more detail in the discussion.

In this paper we present a preliminary set of experiments to determine the kinetic parameters of long and short repeated DNA sequences and to examine possible sequence relationships between the two classes of sequences. Long and short repeated DNAs have been isolated by nuclease digestion and agarose A-50 fractionation. Long repeated DNA has been driven by whole DNA to determine repetition frequency, self-renatured to determine its complexity, and used to drive short repeated DNA to examine cross-hybridization. These criteria show no kinetic difference between long repeated DNA and all repeated sequences in rat DNA and little sequence difference between long and short repeated DNA sequences. Two more sensitive experiments have been used to look for short sequences internal to long repeated DNA fragments. Long repeat DNA has been used to drive varying length whole DNA tracers to determine the interspersion period of the sequences in long DNA in whole DNA. And long DNA fragments have been self-reacted and the fraction of single strands on duplex containing molecules determined by

hydroxyapatite fractionation before and after single strand nuclease digestion. Electron microscope visualization of these renatured duplexes indicate they contain branched structures. All of these data are consistent with the presence of short repeated sequences within long repeated sequences.

Materials and methods

Preparation of DNA

Unlabeled DNA was extracted from rat ascites cells and DNA labeled and extracted from Novikoff hepatoma cells as described in Pearson et. al., (1977b). DNA fragments 3000 to 4000 nucleotides long were prepared by shearing for 45 min at 7500 rpm in a Virtis 60 homogenizer (Britten et. al., 1974) in 0.03 M NaAc pH 6.8. DNA was sheared to 350 nucleotides at 50,000 RPM in 66% glycerol (Britten et. al., 1974). The DNA was then passed over Chelex 100 (Bio-rad) and filtered and precipitated with EtOH.

Sizing DNA fragments

Single stranded DNA fragments lengths were determined by sedimentation through alkaline sucrose gradients. Isokinetic sucrose gradients (Noll, 1967) were formed in Sw41 tubes in 0.1 N NaOH using a Vmix of 10.4 ml, Cflask = 16.0% w/v, Cres = 43% w/v. Gradients were centrifuged from 16 to 24 hours at 40,000 RPM. All tubes contained two makers of known molecular weight and samples were run at least two times. Molecular weights were calculated from sedimentation rates using

the Studier (1965) equations.

Preparation of long repeated DNA fragments

DNA shared to 4000 nucleotides was denatured and incubated at 65 °C in 0.3 M NaCl 0.01 M Pipes pH 6.8 to an equivalent Cot 5 using a factor of 2.31 for the reaction rate increase due to the Na^+ concentration. After incubation, samples were diluted with an equal volume of 0.05 M NaAc 0.2 mM $ZnSO_4$ pH 4.2 and dithithreitol was added to a final concentration of 5mM. The final reaction mix was 0.15 M NaCl, 5mM Pipes, 0.025 M NaAc 0.1 mM $ZnSO_4$ 5 mM DTT pH 4.4.

DNA samples were incubated with S-1 nuclease for 45' at 37 °C and the reaction mix chilled on ice and made 0.12 M PB. Duplex DNA strands were separated by hydroxyapatite chromatography, eluted with 0.5 M PB and chromatographed on Biogel agarose A-50.

The long unlabeled DNA used in the self-reaction and long/short cross-hybridization and interspersion experiments was isolated from 10 mg of 4000 nucleotide DNA. The DNA was denatured for 5 min at 100 °C, incubated to Cot 5 (13 min) and digested with 250 ul of an S-1 nuclease preparation (the gift of Dr. Francine Eden). T
his nuclease preparation
has been extensively characterized (Britten et. al., 1977). The concentration used (25ul/mg) corresponds to a DIG = 0.85 which is 1.7 times the standard incubation. After hydroxyapatite chromatography, 16.8% of the DNA was found duplexed (30 A_{260} units) and this duplex DNA was passed over agarose A-50. 55% of the DNA was excluded.

The ^{3}H labeled DNA used as tracer in the repetition frequency and cross-hybridization experiments was prepared as above. In this preparation, the enzyme to DNA ratio was 25ul/mg, 19% of the DNA was bound to hydroxyapatite and 47% of the repetitive duplexes were excluded from agarose. A third preparation of DNA was used in the melting and renaturation experiments displayed in Figures 7 and 8. 4000 nucleotide DNA was alkaline denatured, incubated to Cot 5 and digested with 25ul/mg of nuclease. The fraction resistant duplex was 14%; 58% of the DNA was excluded from A-50.

Renaturation

Samples which were not to be digested by S-1 nuclease were incubated in 0 .12 M PB at 60°C or in 0.48 M PB at 70 °C. After incubation, samples were frozen in dry ice-ethanol. Samples were thawed and diluted to 0.12 or 0.14 M PB and passed over hydroxyapatite at 60 °C. The fraction bound was eluted after thermal denaturation at 100 °C. The fraction and rate parameters for the renaturation curves were calculated using a non-linear least squares fitting program (Pearson et. al., 1977a).

Melting

DNA samples were melted in 0.12 M PB in a Gilford model 2400 spectrophotometer equipped with a model 2527 thermal cuvette. Samples were melted at a rate of 0.5 °C/min and the A_{260} automatically sampled at 0.5 °C intervals. Hyperchromicity was calculated from the formula

$$H = \frac{A260(98^{\circ}C) - A260(60^{\circ}C)}{A260(98^{\circ}C)}$$

after subtraction of the buffer absorbance at each temperature.

Electron microscopy of DNA

DNA was dialyzed against 0.1 M Tris 0.01 M EDTA pH 8.5, made in 0.05 mg/ml cytochrome c, 40% formamide and spread for microscopy by the modified Kleinschmidt technique of Davis et. al. (1971). The hypophase was 10% formamide in 0.01 M Tris 0.001 M EDTA pH 8.5. Uranyl acetate (5×10^{-5} M in ethanol) was used for staining. The parlodian-coated or carbon-coated 300 mesh grids were rotary shadowed with Pt-Pd (80:20) and observed under a Philips 300 electron microscope.

Results

Figure 1 presents data on the renaturation of 350 nucleotide long rat DNA assayed on hydroxyapatite and the hybridization of a selected repetitive fraction. Table IA is a summary of a least squares fit to these data using a single copy rate determined from the genome size. The repetitive fraction of the genome is overestimated by these measurements beause the duplex fraction bound to hydroxyapatite contains single strand tails. Table IB indicates the true repetitive fraction of the DNA determined by melting and nuclease digestion experiments (Pearson et. al., 1977b).

When 4000 nucleotide DNA is renatured to Cot 5, nucleased and the duplexes sized on agarose A-50, repetitive sequences are fractionated into two classes. A typical column profile is shown in Figure 2. Long (4000 nucleotide) fragments were used to minimize creation of short fragments by random shear and overlap of long repeated sequences. The

Figure 1: Renaturation of 350 nucleotide rat DNA fragments

This data has been collected by a number of investigators (Holmes and Bonner, 1974; Gottesfeld et. al., 1976; Pearson et. al., 1977b). Denatured 350 nucleotide DNA was renatured in .12 M PB at 60 ^{o}C or in .48 M PB at 70 ^{o}C. The fraction single stranded was measured by hydroxyapatite chromatography. Equivalent Cot is the Cot (moles/liter-seconds) times a salt concentration factor (Britten et. al., 1974). The solid lines show the least squares fit (Pearson et. al., 1977a) of the data using a single copy rate fixed at 0.00034 for a genome size of 2.9 pg (Sober, 1968). The actual coefficients for this fit are shown in Table IA.

Also included is the hybridization of a ^{3}H labeled repetitive DNA fraction renaturing before Cot 100 (Pearson et. al., 1977b) driven by whole rat DNA.

(□) Renaturation of whole rat DNA. (●) Hybridization of repetitive rat DNA driven by whole rat DNA.

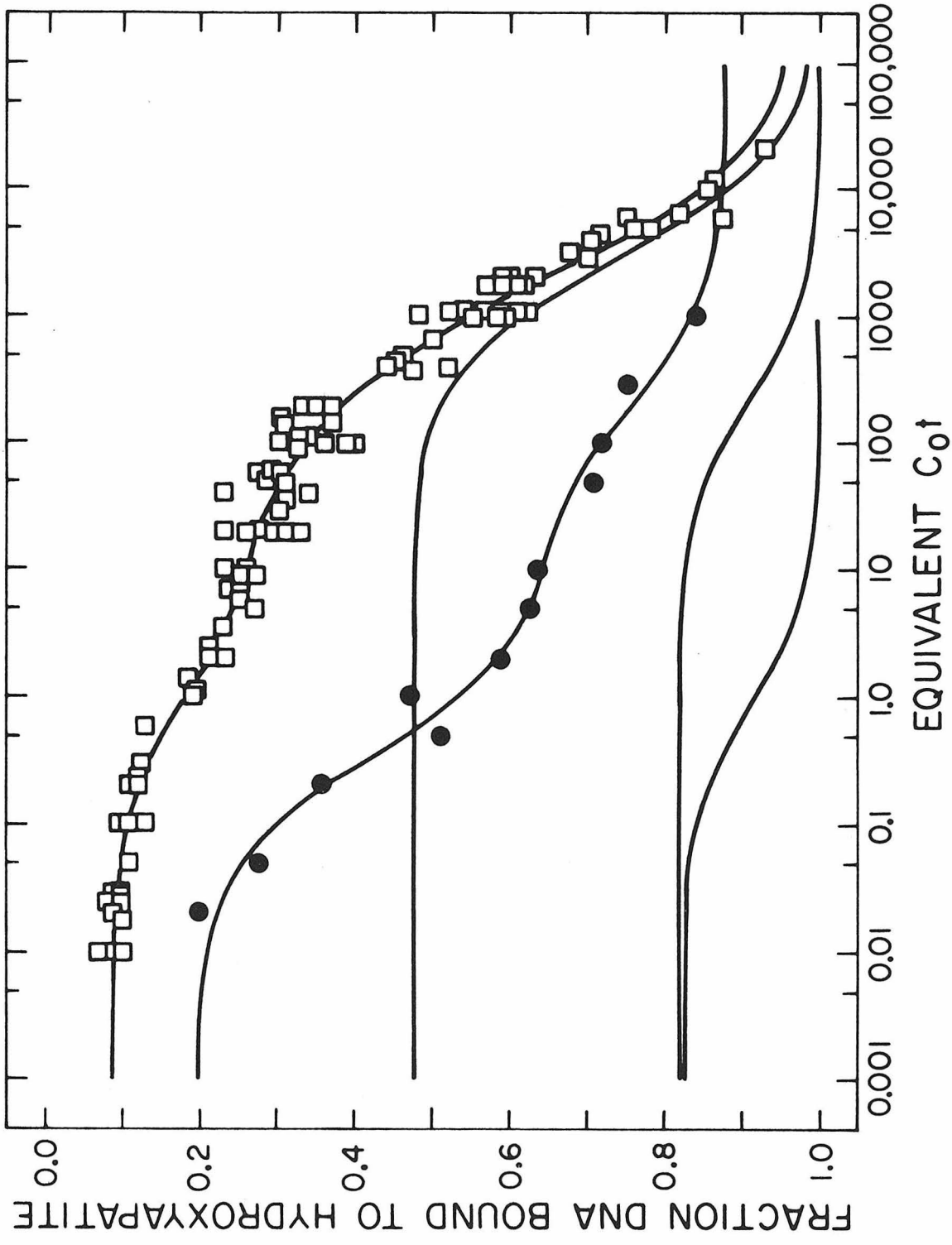

FRACTION DNA BOUND TO HYDROXYAPATITE
0.0
0.2
0.4
0.6
0.8
1.0
EQUIVALENT C_0t
0.001
0.01
0.1
1.0
10
100
1000
10,000
100,000

Table IA

Renaturation of 350 nucleotide Rat DNA fragments

Constrained fit: (single copy rate fixed at 0.00034 for 2.9 pg genome)

Goodness of fit: 3.1%

Component	Fraction	Rate	Cot 1/2	Repetition Frequency
1	0.0882	> 100	0.01	>350,000
2	0.176 (±0.014)	1.10 (±0.39)	0.909	3235
3	0.181 (±0.042)	0.00421 (±0.00211)	238	12
4	0.525 (±0.051)	0.00034	2941	1
Final fraction unreacted:			0.0298 ±0.0194	

Table IB

Sequence distribution in rat DNA

Comp	Fraction of DNA[a]	Size Distribution	Repetition frequency	Complexity (nucleotides)	Number of Sequences
1	0.02[b]		>100,000	530.	
2	0.20[c]	0.4 > 1500 0.6 200-400	3000	1.76×10^5	50 350
3	0.08[d]		12.5	1.69×10^7	
4	0.70	0.75 2500 0.25 >6000	1.	1.85×10^9	560,000 <80,000

Footnotes to Table I

[a] Fraction of DNA is our best estimate of the true fraction duplex. These values differ from those in Table I because of single strand tails bound to hydroxyapatite.

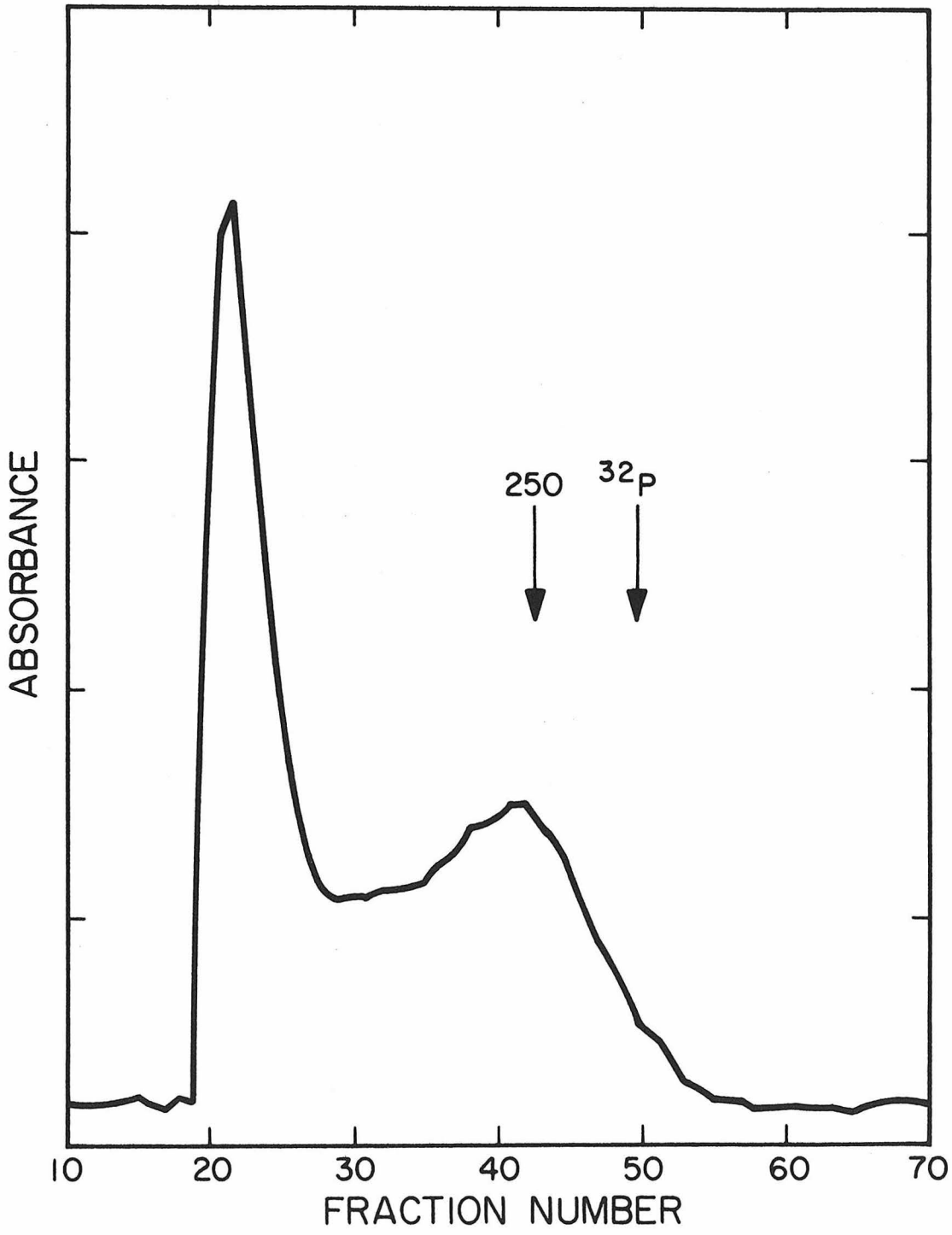
[b] This fraction assumes that .67 of the 0.06 of the duplex bound before Cot 0.05 is 3000 fold repeated sequence (Wu et. al., in prep.)

[c] This fraction includes 0.04 from the binding before Cot 0.05.

[d] This is a rough estimate. None of the nuclease digestion studies separated the low repeat fraction from single copy DNA. From the hydroxyapatite data of Table I, 0.12 of the genome consists of slow repeats. We have assumed 0.67 of that DNA is true duplex.

Figure 2: Profile of rat repeated DNA duplexes on agarose A-50

DNA sheared to 3000 nucleotides was denatured and renatured to Cot 5, digested with S-1 nuclease and bound to hydroxyapatie. The double stranded fraction (15%) was eluted with 0.5 M PB and chromatographed on Bio-gel agarose A-50. The size of the fraction indicated was determined by alkaline sucrose sedimentation.


ABSORBANCE
250
32P
10
20
30
40
50
60
70
FRACTION NUMBER

incubation was carried out to Cot 5 to prevent renaturation of single copy sequences at a higher effective Cot because of the longer fragment length. The long DNA excluded from agarose A-50 is 1500-2000 nucleotides long. The short included DNA is 200-300 nucleotides long.

Hybridization of long repeated DNA sequences

Long labeled repeated DNA fragments were sheared to 350 nucleotides and driven by whole 350 nucleotide long DNA to determine the repetition frequency of the long DNA. Figure 3 shows the hybridization curve for the long DNA. Table II shows the results of a least squares fit to the data. A large fraction of the DNA (60%) hybridizes by Cot 10 with a Cot 1/2 of 0.167. The remaining 20% of the DNA which hybridizes by Cot 1000 contains less frequently repeated sequences from the slower repeat class shown in Figure 1. This part of the reaction may also include some single copy DNA sequences.

The complexity of the long DNA was determined by self-reaction of this preparation. This curve is shown in Figure 4. Again the DNA preparation was sheared to 350 nucleotides to exclude length effects. In this preparation, 12% of the DNA was in duplex at Cot 5 of which 45% (or 5.4% of total rat DNA) was excluded from A-50. Our best estimate for the true fraction repetitive at Cot 5 is 16%, so the long sequences should represent a 1/(.45*.16)=13.9 fold purification of sequences from whole DNA. Table III presents two fits of the data. The first is the best "free" fit with no parameters fixed. In the second fit we assumed the long sequences are a fraction purified from whole repeat DNA. Two components with rates of 13.5 x 2.97 and 13.9 x 0.0404 were fit. The

Figure 3: Hybridization of ^{3}H long repetitive sequences driven by whole rat DNA

^{3}H labeled long repeated DNA fragments were isolated after fractionation on Bio-gel agarose A-50 and sheared to 350 nucleotides. These fragments were driven by a 100-fold excess from Cot 0.01 to Cot 0.1, a 200-500 fold excess from Cot 0.2 to 0.5, a 1000-fold excess from Cot 1.0 to 1000. and a 10,000- fold excess of unlabeled 350 nucleotide whole rat DNA at Cots greater than 1000 in 0.12 M PB at 60 ^{o}C.

(●) fraction single stranded after binding to hydroxyapatite in 0.12 M PB at 60 ^{o}C. The line drawn through the data represents the best least squares fit presented in Table II.

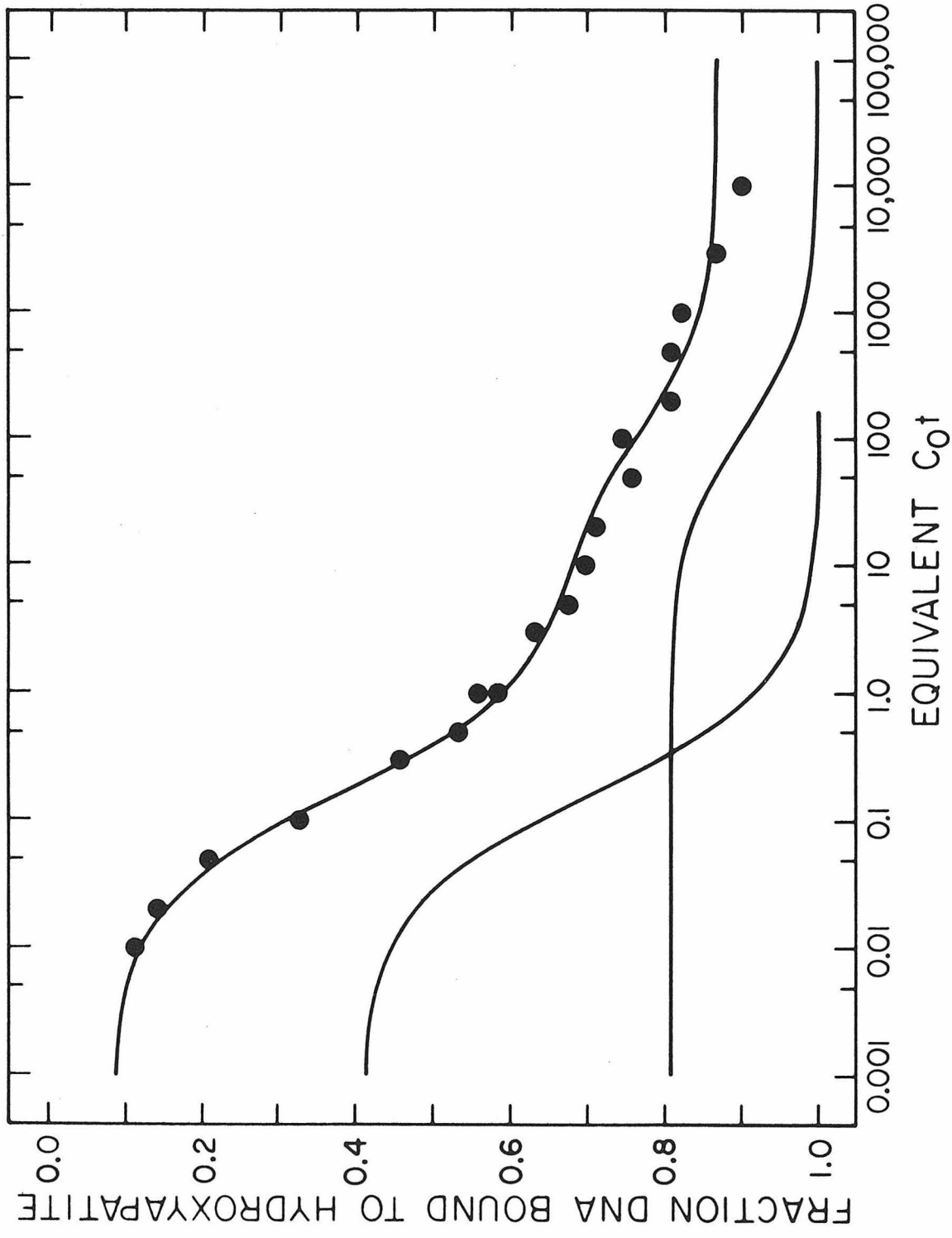

FRACTION DNA BOUND TO HYDROXYAPATITE
0.0
0.2
0.4
0.6
0.8
1.0
0.001
0.01
0.1
1.0
10
100
1000
10,000
100,000
EQUIVALENT C_0t

Table II

Hybridization of sheared long repeated DNA sequences

Driven by whole rat DNA

A. Unconstrained fit: (no parameters fixed)

Goodness of fit: 2.1%

Component	Fraction	Rate	Cot 1/2	Repetition Frequency
1	0.592 (±0.020)	6.0 (±0.86)	0.167	17,600
2	0.173 (±0.017)	0.00286 (±0.00309)	350	10

Final fraction unreacted: 0.132 ±0.014

B. Constrained fit: (slow repetitive and single copy rates set at values from Table I)

Goodness of fit: 2.2%

1	0.600 (±0.020)	5.32	0.188	15,600
2	0.178 (±0.017)	0.00414	242	12
3	0.032 (±0.050)	0.00034	2941	1

Final fraction unreacted: 0.010 ±0.029

Figure 4: Renaturation of sheared long repeated sequence DNA fragments

^{3}H labeled and unlabeled long repeat DNA was isolated as described in Figure 3. The two DNA fractions were mixed and incubated to the Cot values shown.

(●) fraction of ^{3}H labeled counts single stranded after binding to hydroxyapatite. The line through the data is the least squares fit of two components with rates calculated from the rates of the repetitive fractions in the rat times the expected purification of the long DNA components (14-fold).

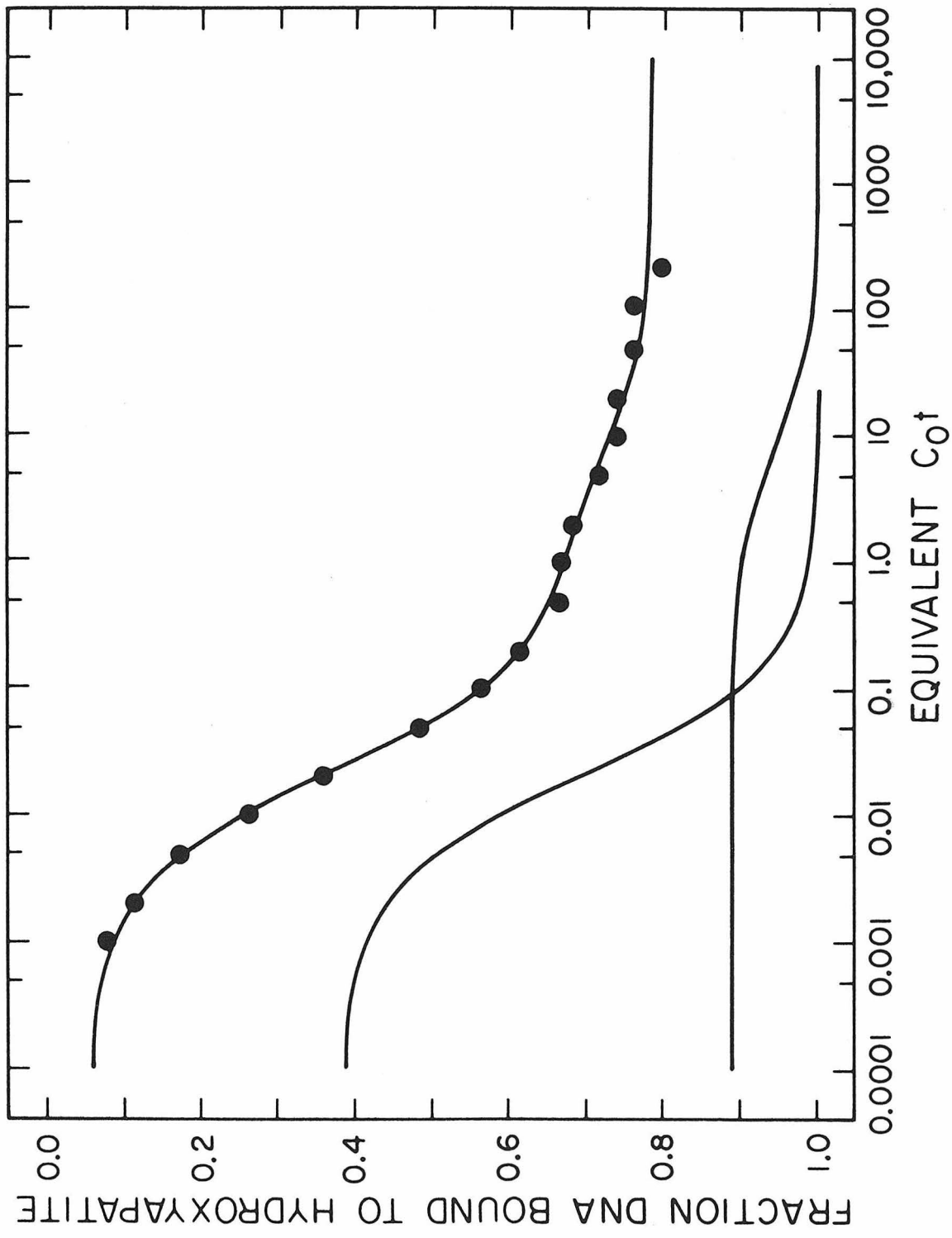

FRACTION DNA BOUND TO HYDROXYAPATITE
0.0
0.2
0.4
0.6
0.8
1.0
0.0001
0.001
0.01
0.1
1.0
10
100
1000
10,000
EQUIVALENT C_0t

Table III

Renaturation of sheared long repeated DNA fragments

A. Unconstrained fit: (no parameters fixed)

Goodness of fit: 0.9%

Component	Fraction	Rate	Cot 1/2	Repetition Frequency
1	0.615 (±0.009)	47.6 (±3.2)	0.021	10,000[a]
2	0.112 (±0.009)	0.114 (±0.040)	8.77	20[a]

Final fraction unreacted: 0.216 ±0.008

B. Constrained fit: (rate constants set to values appropriate for 13.9 fold purification from whole DNA)

Goodness of fit: 1.0%

1	0.618 (±0.085)	0.0242	41.3	10,000[a]
2	0.107 (±0.009)	0.0575	17.4	10[a]

Final fraction unreacted: 0.207 ±0.007

Footnotes to Table III

[a] Repetition frequency of the purified sequences in whole DNA

Figure 5: Cross-hybridization of short repeated DNA sequences driven by long repeated DNA sequences

The unlabeled long repeat fraction used in Figure 4 was used to drive sheared short DNA fragments from the included peak in Figure 2. The driver to tracer ratio was 10 for Cot 0.002, 20 for Cot 0.01 and 100 for Cots greater than 0.5.

(●) fraction of ^{3}H counts single stranded after binding to hydroxyapatite. The line drawn through the data is the best least squares fit of the data with coefficients presented in Table IV.

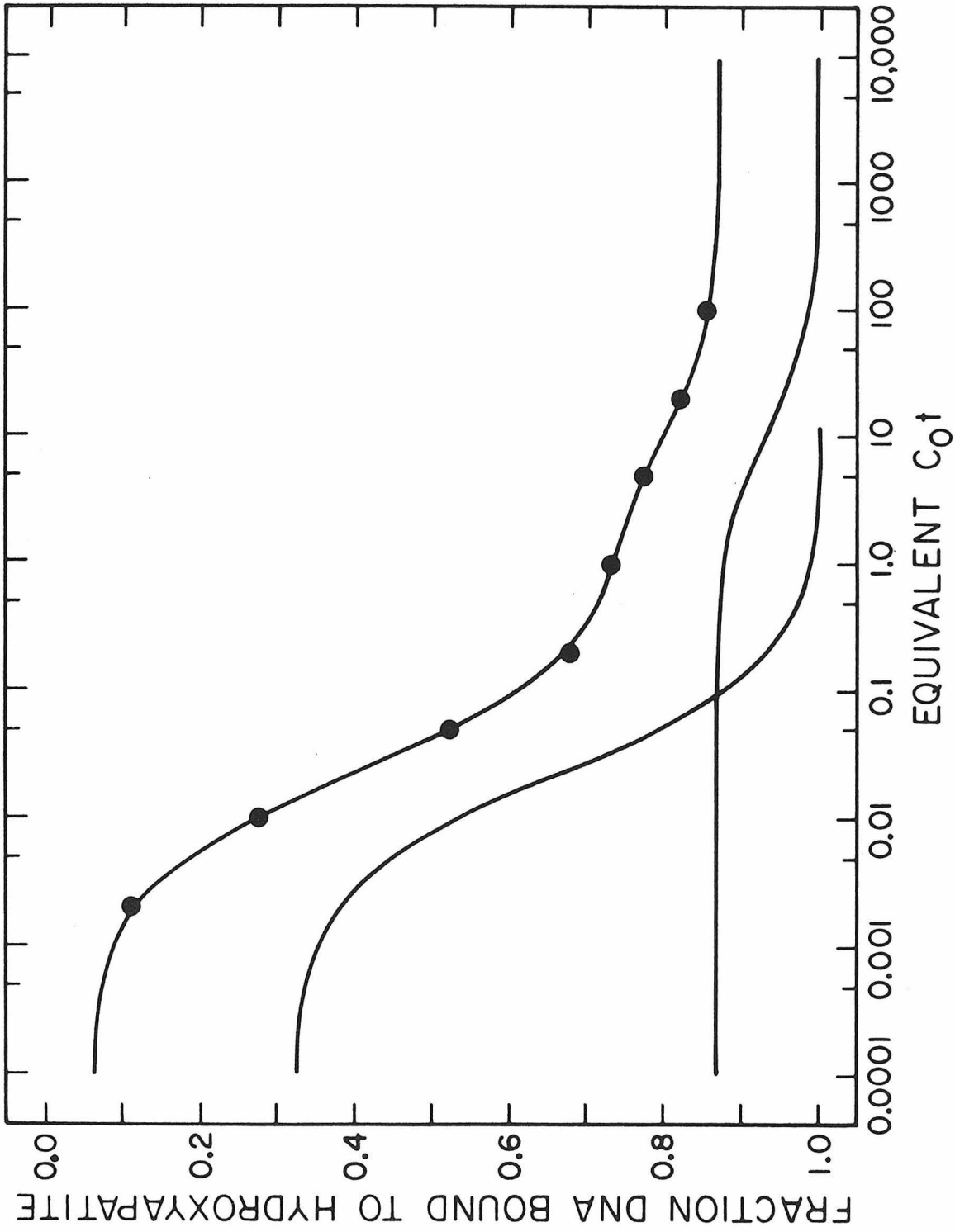

FRACTION DNA BOUND TO HYDROXYAPATITE
0.0
0.2
0.4
0.6
0.8
1.0
0.0001
0.001
0.01
0.1
1.0
10
100
1000
10,000
EQUIVALENT C_0t

Table IV

Hybridization of sheared short repeated DNA sequences

Driven by sheared long repeated DNA sequences

A. Unconstrained fit: (no parameters fixed)

Goodness of fit: 0.6%

Component 1/2 Repetition Frequency	Fraction	Rate	Cot 1/2
1 $10{,}000^a$	0.678 (±0.010)	44.5 (±2.91)	0.0225
2 18^a	0.132 (±0.010)	0.0843 (±0.0282)	11.8

Final fraction unreacted: 0.130 ±0.010

Footnotes to Table IV

[a] This repetition frequency assumes the same 13.9 fold purification factor used in Table III

fits are virtually identical, implying the long DNA may be a pure fraction.

The simplest way to look for cross homology between long and short DNA sequences is to drive one preparation by the other. The results of an experiment in which sheared long repetitive DNA drove sheared short repetitive DNA is shown in Figure 5. The kinetics of the reaction are presented in Table IV. Shorts were not used to drive longs because of the expected cross homology. A large fraction of the short DNA may be derived from long sequences by simple mechanical shear.

Interspersion of long repeated DNA sequences

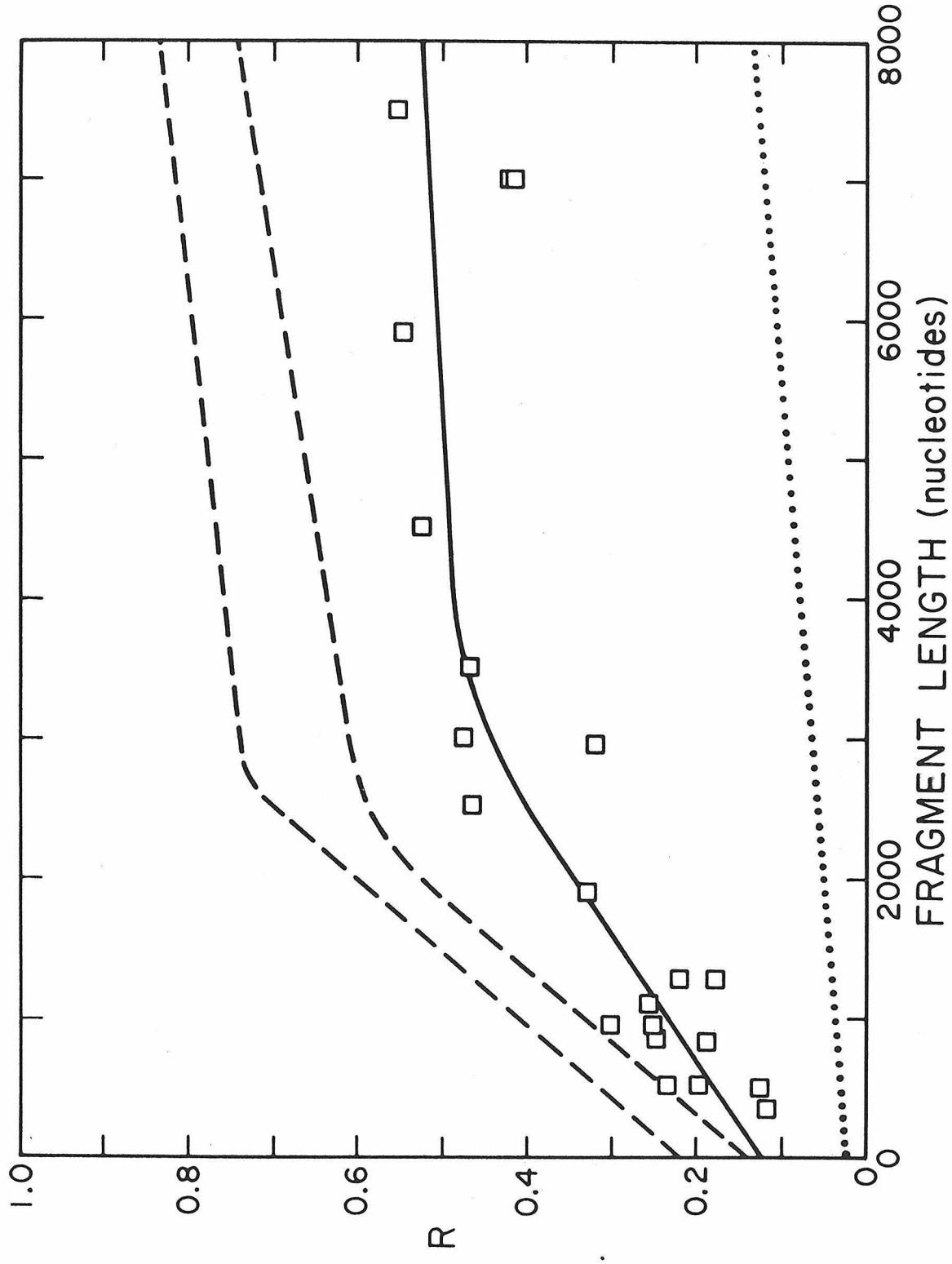
Long repetitive sequences should be relatively free of short repetitive sequence contamination. It is difficult to imagine a mechanical shear, hybridization or nuclease artifact which would create long repeated sequences from short ones. We have used long repeated DNA to drive varying lengths of whole DNA tracers to determine the interspersion period for sequences in long repeated DNA. Figure 6 shows the hybridization at Cot 0.5 of whole DNA tracers as a function of tracer length, corrected for zero time binding in the tracer. This experiment is similar to those done to determine the interspersion period done in other organisms (Davidson et. al., 1973a; Graham et. al., 1974; Angerer et. al., 1975; Schmid and Deininger, 1975; Efstratiadis et. al., 1976; Pearson et. al., 1977b). The dashed lines plotted are interspersion curves which can be drawn through data from DNA tracers driven by whole rat DNA at Cot 5 and Cot 50. Interspersion of long sequences is similar to interspersion of all

Figure 6: The fraction of rat DNA containing long repeated DNA sequences as a function of fragment length.

Labeled whole rat DNA fragments of various sizes were prepared and driven by the sheared long repeat DNA fraction used in Figures 3 and 4. The fraction of the fragments (F) containing duplexes was measured by binding to hydroxyapatite. The value plotted (R) is the value of (F) corrected for the amount of zero-time binding in the long tracers (Z). The values plotted are:

$$R = \frac{F - Z}{1.0 - Z}$$

The values of (Z) were calculated from a linear fit of the zero-time binding data from Pearson et. al. (1977b).


1.0
0.8
0.6
0.4
0.2
0
R
0
2000
4000
6000
8000
FRAGMENT LENGTH (nucleotides)

repeated DNA sequences in the rat. The increase in hybridization from 12% at zero length to 34% at 1800 nucleotides cannot be due to reaction of 1500 nucleotide sequences. The difference in the single strand tail length for a 1500 nucleotide sequence would only account for a 20% change from 12% to 14.4% if only long sequences had hybridized. Some of the sequences in the long repeated DNA preparation must be able to hybridize with short sequences interspersed throughout the genome.

The 3200 nucleotide "best fit" interspersion period is slightly longer than the 2500 nucleotide values measured in whole rat DNA. This is an encouraging result. If some of the long sequences were not shared by short repeated sequences, the interspersion period of those sequences would be longer.

Self-reaction of unsheared long repeats

If a long repetitive sequence contains individual internal short repeat sequences, self-reaction of unsheared long repeats should generate short duplex regions with single strand tails containing other repeated sequences. We have visualized long and short repeated DNA fragments in the electron microscope after nuclease digestion and A-50 fractionation. While most of the molecules are linear, a number of more complex circular and branched structures are found (Plate IA). The structures displayed are more than 95% double stranded as measured by optical hyperchomicity. Thus some of the structures must be due to sequence arrangements internal to the long repeat sequences.

Plate IA: Long repetitive DNA fragments

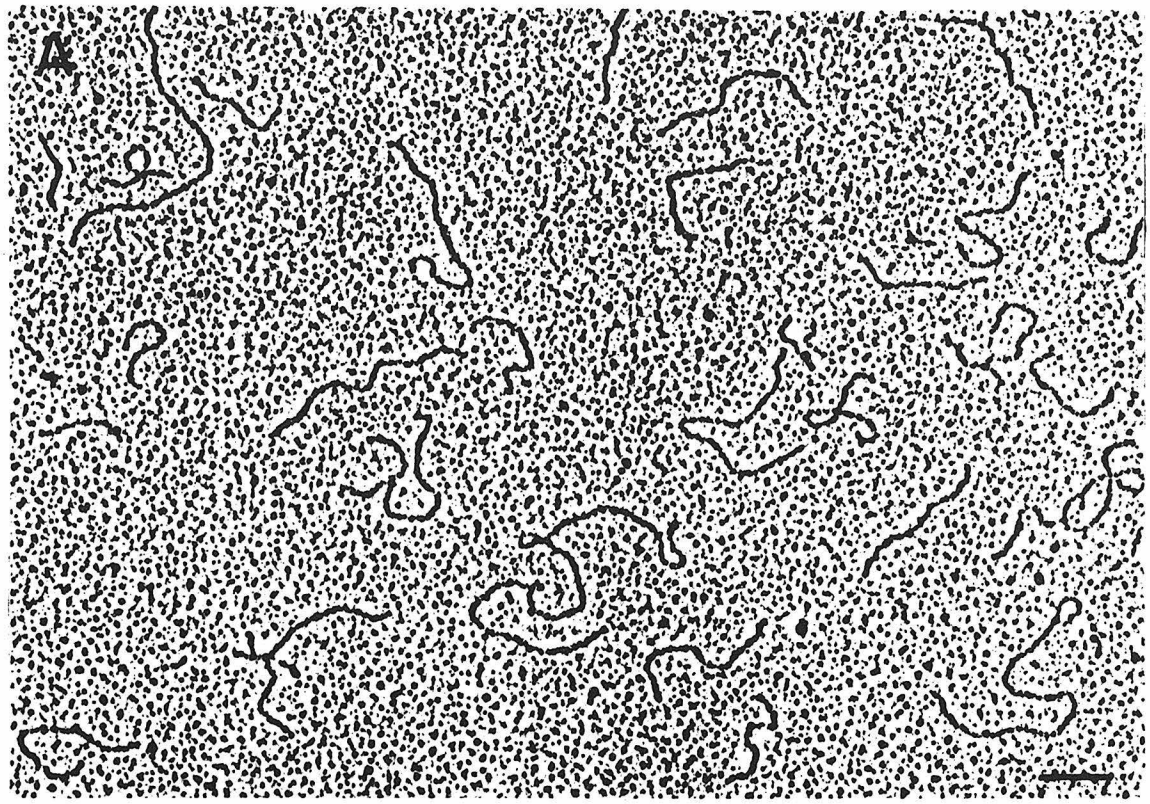
4000 nucleotide DNA was denatured and incubated to Cot 5, treated with S-1 nuclease and the double strand fraction bound to hydroxyapatite. This double strand fraction was chromatographed on agarose A-50 and an excluded fraction (approx. fraction 20 in Figure 2) spread for electron microscopy.

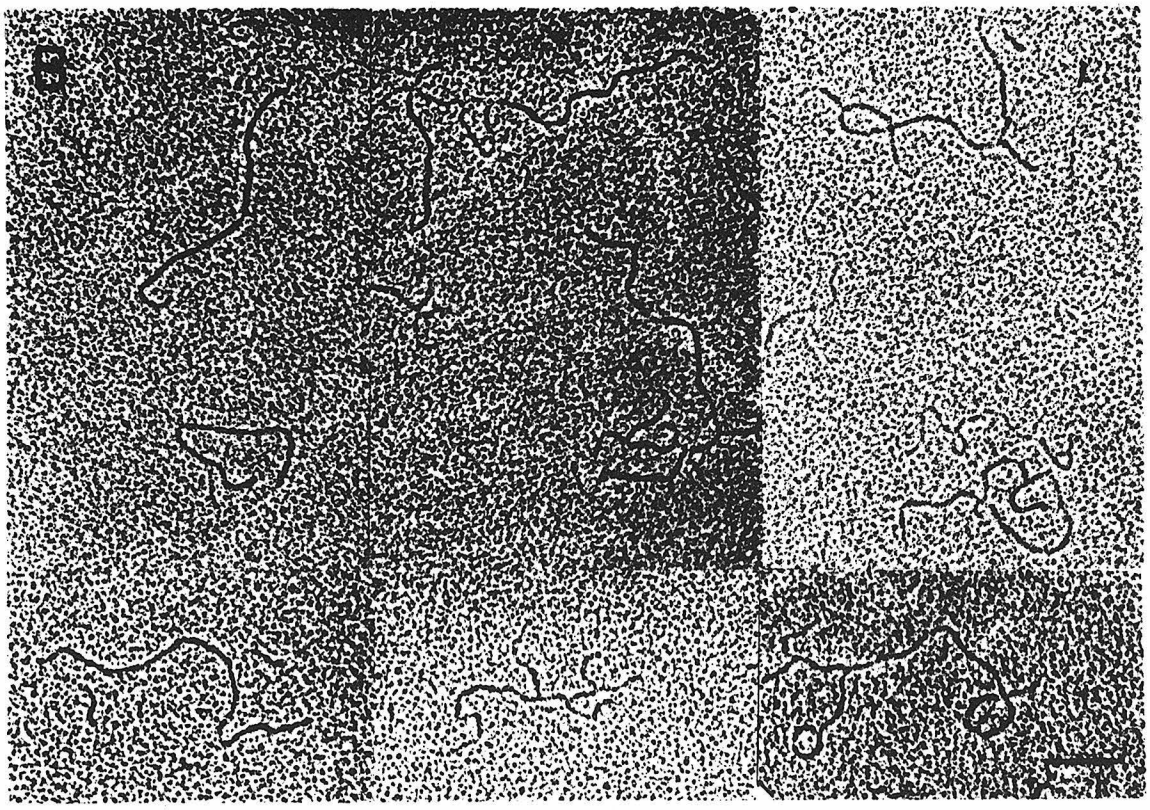
The bar in the lower right of the photograph is 0.2 microns (about 600 nucleotides). The number average length of the excluded fragments is 2400 nucleotides.

Plate IB: Twice renatured long repetitive DNA fragments

Long repeated DNA fragments isolated as in Plate IA were denatured, allowed to renature a second time to Cot 5 and spread for electron microscopy. The bar in the lower right is 0.2 microns.

The number average length of the structures formed was 2000 nucleotides in a range from 750 to 5000 nucleotides.


A


B

Long repeat fragments can be denatured, renatured and then visualized by microscopy or assayed for single strand tails by hydroxyapatite fractionation before and after S-1 nuclease digestion. When isolated long fragments were denatured, renatured to Cot 5 a second time and melted, the duplexes showed 80% hyperchromicity (Figure 7). A native DNA standard run in parallel is also shown. The twice renatured material has more than 80% of the hyperchromicity of native DNA with a T_m of 81 ^{o}C. This may not indicate significant single strand regions; the slight decrease in hyperchromicity may be due to impurities introduced during the preparation and renaturation of the DNA.

This twice renatured DNA was also spread for microscopy using the Kleinschmidt technique. Plate IB displays some of the structures formed. A large fraction of these structures are not linear and contain branches and tails. This suggests that when the long sequences renature a second time internal repeat duplexes may form. It is not surprising that the molecules melt as if they were all duplex; a second renaturation to Cot 5 is equivalent to Cot 50 because of the decreased complexity of the isolated long fraction. The branches seen in the micrographs must also be at least 80% duplex and may reflect hybridization of one ordering of short sequences with a different ordering of short sequences on other strands.

To look for single strands formed by renaturation of internal short repeats, long repeated sequences were renatured through a range of Cots from 0.0005 to 10 and the samples divided. Half of each sample was treated with S-1 nuclease to digest single strands and the duplex regions were measured on hydroxyapatite. The other half was put

Figure 7: Melt of twice renatured long repeated DNA fragments

The long repeat DNA fraction isolated and used in Plate I was denatured and incubated to Cot 5 at 60 °C in 0.12 M PB a second time. The twice renatured material was also used for Plate IB. This sample, a sample of the original long repeated fraction (incubated to Cot 5 once during isolation) and native DNA were melted in a spectrophotometer equipped with a thermal cuvette.

(□) long repeated DNA fragments twice renatured to Cot 5. (×) long repeated DNA fragments renatured once to Cot 5. (◇) native DNA.

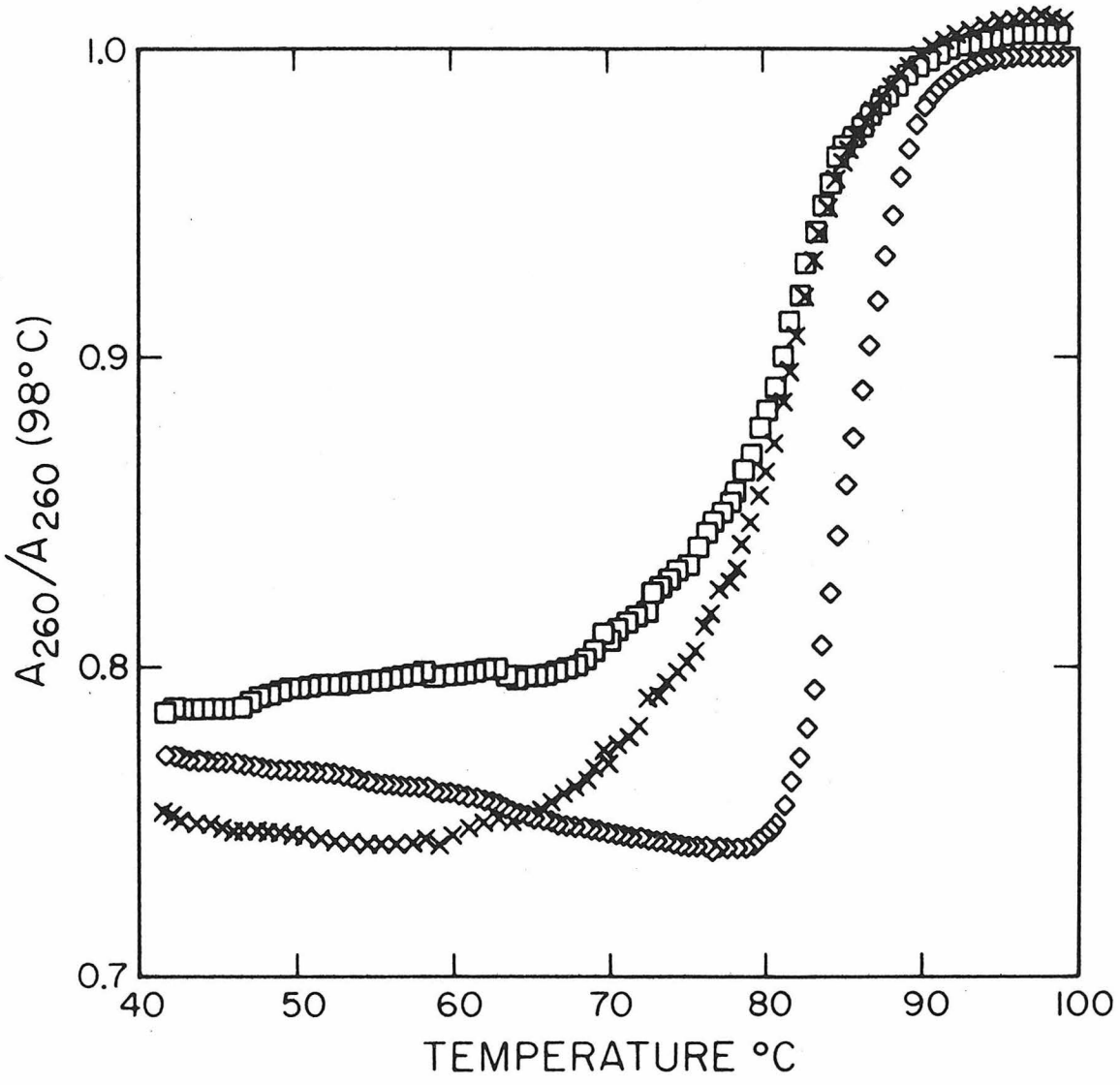
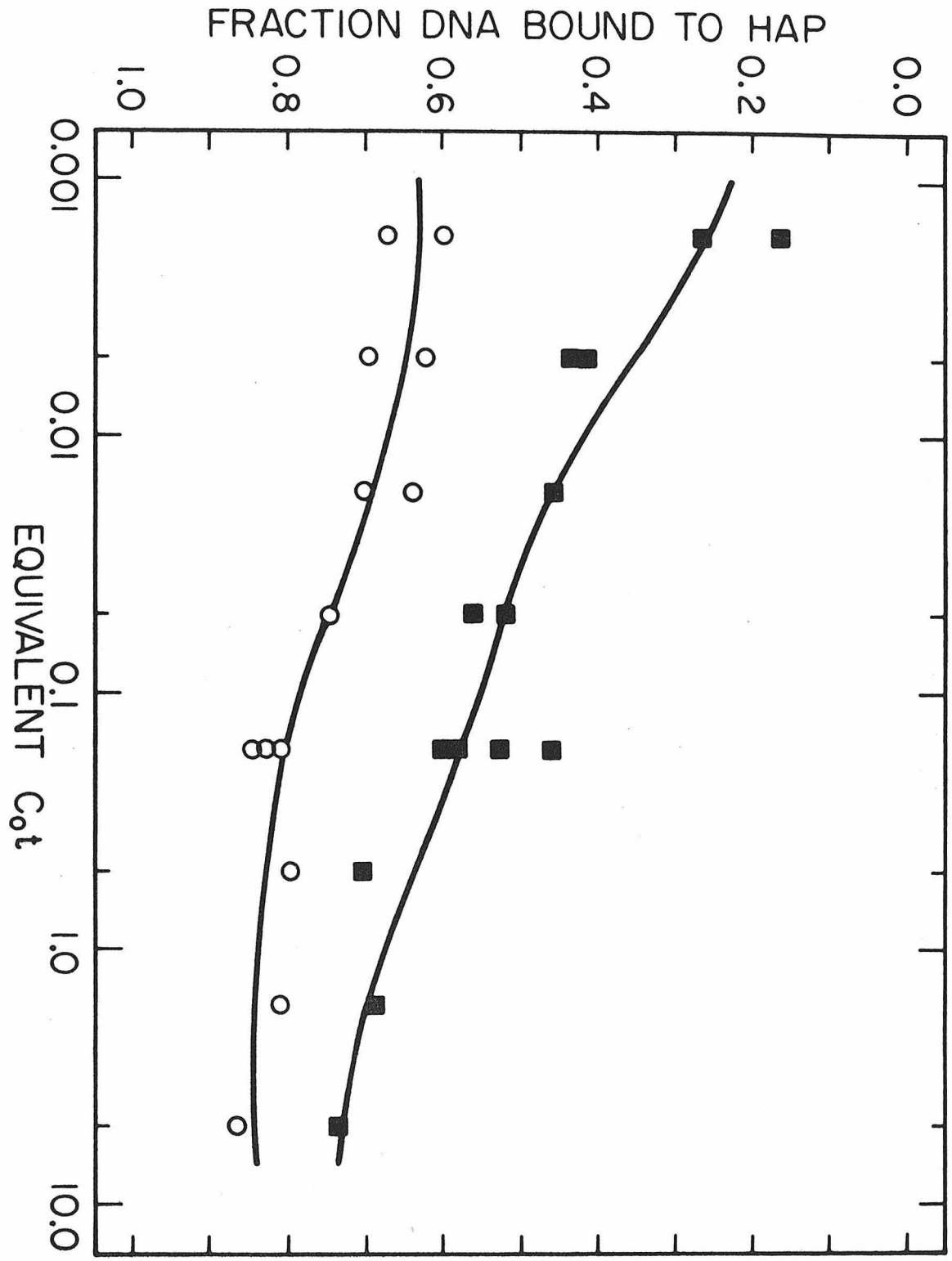

1.0
0.9
0.8
0.7
A_{260}/A_{260} (98°C)
40
50
60
70
80
90
100
TEMPERATURE °C

Figure 8: Renaturation of unsheared long repeated DNA sequences

A long ^{3}H labeled DNA fraction was isolated as described in Materials and Methods. The DNA was alkaline denatured and incubated to the Equivalent Cot values shown in 0.3 M NaCl 0.01 M Pipes pH 6.7. The samples were then divided. One half was treated with S-1 nuclease (50 ul/mg), made 0.12 M PB and the duplex fractions separated on hydroxyapatite. The other half of the sample was not treated with nuclease but made 0.12 M PB and duplex containing strands measured on hydroxyapatite.

(■) Fraction DNA bound to hydroxyapatite after S-1 nuclease. (O) Fraction DNA containing duplex regions (binding without nuclease).


FRACTION DNA BOUND TO HAP
1.0
0.8
0.6
0.4
0.2
0.0
0.001
0.01
0.1
1.0
10.0
EQUIVALENT C_0t

directly over hydroxyapatite and the duplex containing fraction with any single strand tails bound and eluted. The renaturation curves with and without nuclease are shown in Figure 8. The differences between the digested and undigested samples show 50% or more of the hybrid molecules are single stranded. This may be due to internal short sequences or random overlap of fragments which are shorter than their parent sequences. These problems are discussed further below.

Discussion

The measurements we have made on the repetition frequency and complexity of long repetitive DNA in the rat show little difference between the long repeated sequences and repeats in whole DNA. The repetition frequency of long sequences is quite close to the repetition frequency of repeats in whole DNA and the complexity of the long size class is exactly the complexity expected for a fraction of sequences purified from whole DNA. If long and short repeated DNA sequences share the same sequences, the 7.0% of the DNA isolated as long repeated DNA would have 16% of rat DNA complexity and hybridize two fold more slowly. This experiment is not very sensitive, however, and inconsistency with the evidence from the other experiments may be more apparent than real. The least squares fit provides the interpretation that the data are consistent with long repeats being a pure fraction. The presence of two components and the variable recovery of the slow component provide a lot of flexibility in the interpretation. The slow repeat component will not be fully recovered by hybridization to Cot 5, and its presence allows the fitting program to adjust the fraction of the slow component to fill in renaturation by some high complexity component.

The question of sequences shared by long and short repeats is directly addressed by cross-hybridization of short repetitives driven by long repetitive sequences. Here the result is different; the simplest interpretation is that these two classes share sequences. This is not surprising, many of the short sequences must be derived from randomly sheared long sequences which formed short complements with other sheared long sequences.

Measuring the interspersion period of sequences in long repeated DNA is a more accurate probe for short/long sequence overlap. Our preparation of long repeats hybridizes with short sequences which are interspersed like all repeated sequences in whole DNA. At the short interspersion period (2500 nucleotides), 42% of the DNA is bound when hybridized to Cot 0.5. The lower Cot was used to adjust for the ten fold lower complexity of the repetitive driver. The result suggests that 42%/57%=74% of repeated sequences interspersed with a period of 2500 nucleotides hybridizing by Cot 5 in rat and 42%/73%=58% of sequences hybridizing by Cot 50 can be driven by our long repeat preparation. The best fit data from this experiment suggest that some sequences in long repeated DNA are interspersed with a longer period.

The self-reaction of long repeat sequences presents a similar result. The initial self-reaction of long repeat fragments indicates that more than 50% of the structures containing duplex are single stranded. Later in the reaction there are no single stranded regions and the structures formed are branched.

All of these results suggest that long repeated DNA fragments may contain internal sequences which hybridize with short sequences interspersed throughout the genome. Unfortunately there are substantial problems with each of the experiments. The cross-reaction and interspersion experiments are sensitive to cross contamination of shorts with longs and longs with shorts. And the self-reaction result may be artifactual due to the unknown length of the internal elements of the long repeat sequence.

The kinetic complexity and cross reaction experiments are extremely sensitive to contamination of short repeated DNA with long repeated sequences found in short fragments. Random shear in the initial preparation of the DNA or mechanical shear of the repeat duplexes on hydroxyapatite generate short sequences from long ones. Melting experiments (Davidson et. al., 1973b; Goldberg et. al., 1975; Pearson et. al., 1977b) suggest many of the short repeat sequences are different from long repeats. Short sequences exhibit more mismatch than long sequences. But some short sequences must be derived from long sequences and the melt of the short repetitive fragments shows a high precision component (Pearson et. al., 1977b).

It is difficult to estimate how many of the short sequences are derived from long sequences without knowing the <u>in vivo</u> length of long sequences in the genome. We do not know that length. It may be longer than the 4000 nucleotide fragments used in these experiments. The length of the long duplexes excluded from agarose A-50 is about 1500-2000 nucleotides, but these may include molecules which were mechanically sheared during hydroxyapatite fractionation.

It is possible to put an upper limit on the amount of contamination by placing a lower limit on the number of short sequences interspersed throughout the genome. At Cot 5, 57% of 2500 nucleotide DNA fragments contain a short repeated DNA sequence. If the short repeated sequences are 250 nucleotides long, 250/2500 = 10% of the bound DNA is repetitive so 5.7% of the DNA must contain short repetitive sequences hybridizing by Cot 5. At this Cot, .16 of the genome is in true duplex so 0.057/0.16=0.356 of duplex sequences must be short. We find 0.5 to 0.6 of duplex DNA included on agarose A-50, so 15% to 25% of that DNA or 15%/50%=30% to 25%/60%=42% of the short duplexes may be derived from long sequences. This is an upper limit. Many short sequences are much longer than 250 nucleotides and others may be interspersed at a longer period. If 33% of short sequences are derived from long sequences, short sequences would drive all long sequences in a cross reaction experiment. Conversely, 30%-40% of the short fragments would be driven by long sequences.

In addition, the long sequences may be contaminated by short sequences. This could account for more of the long/short cross-hybridization reaction and the apparent short interspersion period of sequences in long DNA. The hybridization curves of Figure 3 indicate that as much as 20% of the long sequences may contain single copy DNA (10% reaction at Cot 2900 and 10% unreacted). In that case, 9.6% (16% duplex at Cot 5 * 60% included on A-50) of the 20% single copy or 1.92% could be short repeat contamination. In addition, an equal fraction might be present because of incomplete digestion of single strand tails. This 4% contamination would drive the short DNA sequences at a 25 fold reduced rate. This reaction, following a conservative 20%-30% cross

reaction due to short contamination by longs might account for the results of figure 4. The problem would be more serious if as much as 10% of the single strand tails were not digested by the nuclease. Such contamination would be difficult to detect by melting experiments, which are the only independent test of long sequence purity. A low melting shoulder on the long repeated DNA melting curve could be due to short repeated DNA contamination or mismatched long repeated sequences.

Some of the interspersion of long repeat sequences in whole DNA is due to sequences spaced at lengths longer than the average 2500 nucleotide interspersion period in rat (Pearson et. al., 1977b). If short DNA sequences were spaced at 1000 nucleotides and long repeated sequences were spaced at 4000 nucleotides, a small (4%) amount of short sequence overlap or contamination might explain the results. These data, and data from these kinds of experiments in general, is not accurate enough to determine the fraction of short sequences required to explain the reaction. In addition, much of the interspersion curve may be caused by sequences which renature well before Cot 5. The 4% contaminant of short repeated sequences in long fragments may be more repetitive that whole repeated DNA. This is likely, more repeated sequences will have better chances for multiple collisions to form long concatenate structures which could be nuclease resistant. If so, the contaminant could account for some of the binding at the short fragment lengths. This experiment does not measure the repetition frequency of the sequences driving the sequences generating the interspersion curve.

The concentration/repetition frequency argument affects the interpretation of much of these data. One can never be certain that a very small level of contaminant repeated with higher than average frequency is not the cause of the hybridization. This problem particularly affects the cross-hybridization and interspersion experiments.

Interpretation of the self-renaturation of long repeated sequences shown in Figure 8 is dependent on the length of the long sequences in the DNA. Figure 9 shows two possible interpretations. If the long repeated sequences are longer than the fragments of DNA renatured (Figure 9A) the second time, 50% of the DNA on a duplexed structure bound to hydroxyapatite will be single stranded (Britten and Davidson, 1976) because of random overlap. On the other hand, the single strands associated with duplexes may be due to more complex branched structures as shown in Figure 9B. Electron microscope visualization of the long self-reaction products supports the second model. The multi-ended structures in Plate IB could reflect different internal sequence arrangements in different long repetitive fragments. They may also be due to overreaction of the repeated sequences allowing single strand tails to react with other molecules to build up larger structures with branches.

Problems with the long fragment self-reaction experiment are more easily solved. These results are not affected by moderate levels of contamination in the long sequence population as these experiments look at the majority of the mass of the long sequences. Self-driven experiments measure properties of the most prevalent species.

Figure 9: Formation of duplexes between long repeated DNA fragments: two models

This figure presents two models for the result of Figure 7: that renatured long repetitive DNA fragments form single strand tails on a second renaturation. Part A shows what would happen if the fragment length L were shorter than the repetitive sequence length R. Part B shows the sequence arrangement necessary to explain single strand structures if the fragment length L is greater than the repetitive sequence length R.

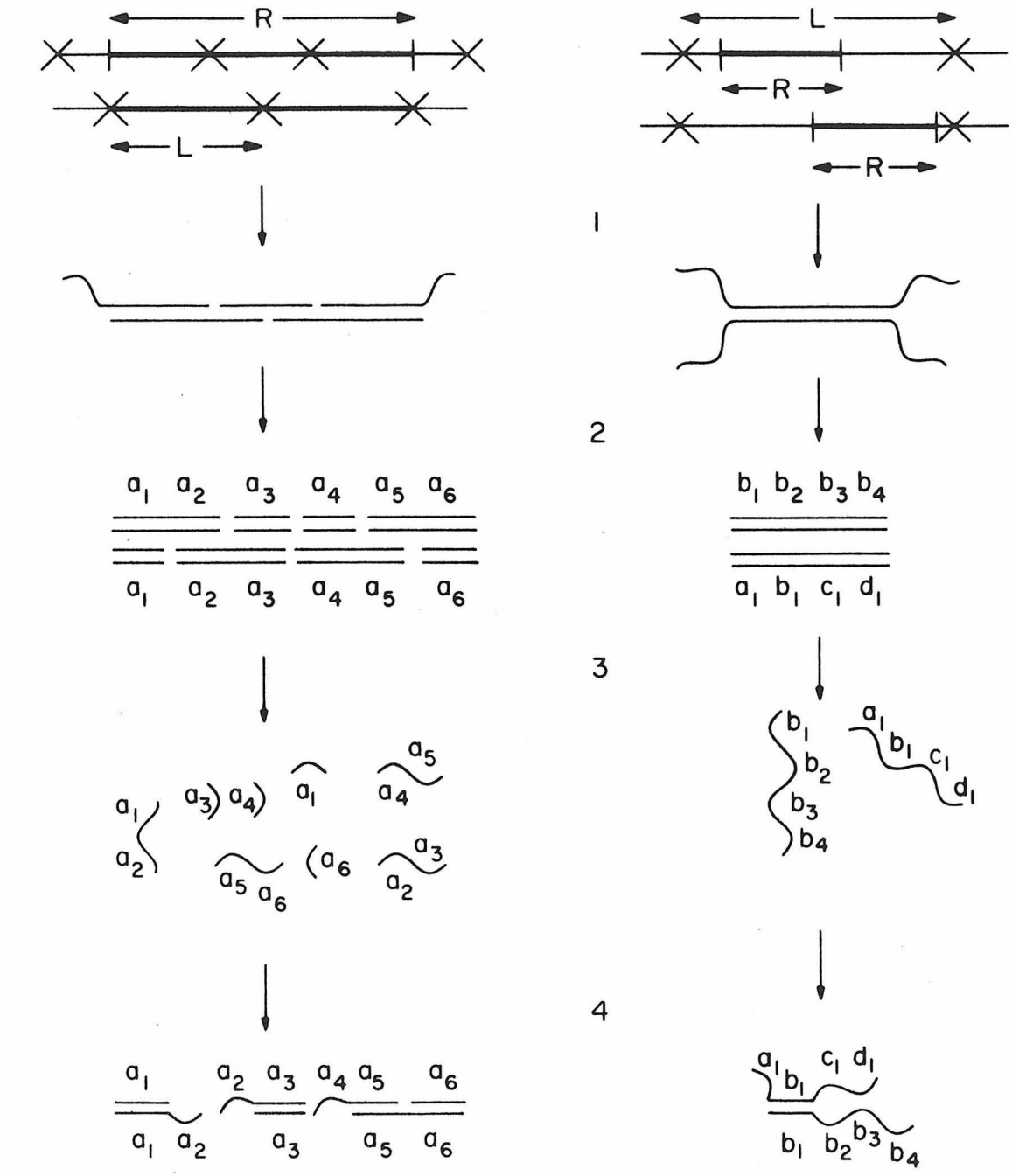
The numbers (1 - 4) are the steps in the preparation of long repeated DNA fragments from randomly sheared (X marks random cuts) cDNA fragments.

1. Denaturation and renaturation of fragments of length L.

2. S-1 nuclease digestion of renatured duplexes. The gaps between a_1a_2 and a_3 for example are due to the single strand breaks in the parent duplex.

3. Second denaturation of long repetitive fragments including molecules from other regions of the genome.

4. Second renaturation of long repeated fragments. The extent of single strand regions depends on the number of collisions for the case of Part A while more collisions among Part B structures would lead to more complex molecules with single stranded regions.


A.
B.
R
L
L
R
R
1
2
3
4
a1 a2 a3 a4 a5 a6
a1 a2 a3 a4 a5 a6
b1 b2 b3 b4
a1 b1 c1 d1
a1 a2 a3 a4 a5 a6
a1 a2 a3 a4 a5 a6
b1 b2 b3 b4
a1 b1 c1 d1
a1 b1 c1 d1
b1 b2 b3 b4

Longer DNA can be used to make repetitive duplexes. If duplexes from longer fragments are not longer than long duplex structures from shorter fragments, the length of the long repeats has been found. Full sequence length fragments will not have single strand tails due to overlap errors. Hydroxyapatite binding with and without nuclease digestion may conclusively show short internal repeat sequences. These short duplexes may also be sized after S-1 nuclease to find the length of the short sequence element.

Electron microscopy can also be used to look for internal short repeats. At early stages of the self renaturation each molecule should only have experienced one collision on average. If the long repeats are single stranded because of random overlap, the hybrid molecules should be linear with only two tails. If internal sequences are hybridizing the molecules should have 3 and 4 tail structures due to single strands on either side of the duplex. This topological criterion can distinguish between the two possibilities and is independent of the fragment length of the strands used.

Conclusions

The renaturation, cross-hybridization and melting experiments we have presented all suggest that there are no significant differences in complexity or repetition frequency between long and short repeated DNA sequences in the rat. Eden et. al. (1977) have measured the repetition frequency, complexity and sequence overlap between long and short repeated sequences in the sea urchin. They found that the repetition frequencies of the long and short repetitive sea urchin

sequences in whole DNA are similar, with the long sequences repeated slightly less. And they found the kinetic complexity of the short sequences is about three times that of the long DNA sequences, reflecting the fraction of long sequences in the whole sea urchin repetitive sequence population (about 30%).

The cross-hybridization results - using long repetitive sequences to drive short repetitive sequences - in the sea urchin are quite different from the results with rat DNA. The short repetitive tracer was driven at a rate 1/10 that of the long repetitive driver. This result suggests that 10% or fewer of the sequences in long repeated DNA fragments can hybridize with short repeated DNA sequences. The result contrasts sharply with the results presented for the rat - we cannot show any difference between short repetitive tracer hybridization and the self-reaction of the long repetitive driver.

We believe there are two explanations which could account for the conflicting results. First, the higher fraction of long repeated sequences in the rat may cause more difficult contamination problems. Second, the internal structure of the long repeated duplexes may be different in the rat.

Under virtually identical digestion and fractionation conditions, we find from 40% to more than 50% of rat repeated DNA sequences from 4000 nucleotide fragments are excluded on agarose A-50 while Eden et. al. (1977) found about 30% of the sequences from 2000 nucleotide fragments were excluded. The calculations we presented earlier suggest that as much as 70% of rat repeated DNA may be longer than 1000 nucleotides. Experiments using a wide range of digestion conditions for

sea urchin DNA (Britten et. al., 1977) show that 47% of isolated repeat sequence duplexes are longer than 1000 nucleotides after mild digestion. A much smaller fraction of true short repeats in the rat genome would make isolation of a short repeat fraction free of long repeat contamination much more difficult. Much of the discussion described the technical problems involved in these kinds of experiments; while it may be simple to show two repetitive sequence populations are distinct, it is difficult to unambiguously demonstrate true sequence overlap.

Some of the long repeated sequences must include the transcribed multigene families (Galau et. al., 1976) and must be distinct from short repeated sequences. Long repeated DNA sequences also form higher precision hybrids than most of the short repeated DNA (Davidson et. al., 1973b; Goldberg et. al., 1975, Britten et. al., 1977; Pearson et. al., 1977b) which are probably distinct from short sequence hybrids.

We believe there is evidence that long and short repeated DNA fragments share sequences in the rat. Possible contaminant artifacts contribute to ambiguity in many of our results, but we have not been able to provide any strong evidence that long and short repetitive sequence sets are distinct in the rat. The circular structures found in Plate I may be due to tandem arrangement of short sequences in long DNA fragments. The branched structures must reflect different arrangements of sequences internal to long repeated DNA fragments. This interpretation is consistent with our most sensitive tests for long/short sequence sharing: the cross-hybridization, interspersion and unsheared self-renaturation experiments.

If long and short repetitive sequences are shared in the rat a number of structural and organizational questions are raised. Some short repeated sequences may be present as a single sequence within a long repeated sequence or long repeated sequences may be arranged as tandem arrays of short repeats. This interpretation is suggested by the cross-hybrization and interspersion experiments.

In addition, some long repeated sequences may be made up of a number of different repeated sequences which are also found in the genome as short interspersed repeat sequences. If the same short sequences were arranged differently in different long sequences, the results of the electron microscopy and long self-renaturation experiments are easily explained. Either structural model could provide the "integrator gene" function suggested by Britten and Davidson (1969).

References cited

Angerer, R. C., Davidson, E. H. and Britten, R. J. (1975) "DNA sequence organization in the mollusc Aplysia californica" Cell 6:29-39

Bonner, J., Garrard, W. T., Gottesfeld, J. M., Holmes, D. S., Sevall, J. S. and Wilkes, M. M. (1974) "Functional organization of the mammalian genome" Cold Spring Harb. Symp. Quant. Biol. 38:303-310

Birnstiel, M., Telford, J., Weinberg, E. and Stafford, D. (1974) "Isolation and some properties of the genes coding for the histone proteins" Proc. Nat. Acad. Sci. US 71:2900-2904

Britten, R. J. and Davidson, E. H. (1969) "Gene regulation in higher cells: A theory" Science 165:349-357

Britten, R. J. and Davidson, E. H. (1976) "Studies on nucleic acid reassociation kinetics: empirical equations describing DNA reassociation" Proc. Nat. Acad. Sci. US 73:415-419

Britten, R. J., Graham, D. E. and Neufeld, B. R. (1974) "Analysis of repeating DNA sequences by reassociation" Methods in Enzymology (L. Grossman and K. Moldave, eds.) Vol 29 part E 363-418

Britten, R. J., Graham, D. E., Eden, F. C., Painchaud, D. M. and Davidson, E. H. (1977) "Evolutionary divergence and length of repetitive sequences in sea urchin DNA" J. Mol. Evol. in press

Brown, D. D. and Sugimoto, K. (1974) "The structure and evolution of ribosomal and 5S DNAs in Xenopus laevis and Xenopus mulleri: the evolution of a gene family" Cold Spring Harb. Symp. Quant. Biol. 38:501-505

Chamberlin, M. E., Britten, R. J. and Davidson, E. H. (1975) "Sequence organization in Xenopus studied by electron microscopy" J. Mol. Biol. 96:317-333

Davidson, E. H. and Britten, R. J. (1973) "Organization, transcription and regulation in the animal genome" Quart. Rev. Biol. 48:565-613

Davidson, E. H., Hough, B. R., Amenson, C. S. and Britten, R. J. (1973) "General interspersion of repetitive with non-repetitive sequence elements in the DNA of Xenopus" J. Mol. Biol. 77:1-23

Davidson, E. H., Graham, D. E., Neufeld, B. R., Chamberlin, M. E., Amenson, C. S., Hough, B. R. and Britten, R. J. (1974) "Arrangement and characterization of repetitive sequence elements in animal DNAs" Cold Spring Harb. Symp. Quant. Biol. 38:295-301

Davis, R. W., Simon, M. and Davidson, N. (1971) "Electron microscope heteroduplex methods for mapping regions of base sequence homology in nucleic acids" In Methods in Enzymology (L. Grossman and K. Moldave, eds.) Vol XXI, Part D p. 413-428

Eden, F. E. Graham, D. E., Painchaud, D. M., Davidson, E. H. and Britten, R. J. (1977) "Exploration of long and short repetitive sequence relationships in the sea urchin genome" Nuc. Acids Res. in

press

Efstratiadis, A., Crain, W. R., Britten, R. J., Davidson, E. H. and Kafatos, F. (1976) "DNA sequence organization in the lepidopteran Antherea pernyi" Proc. Nat. Acad. Sci. US 73:2289-2293

Galau, G. A., Chamberlin, M. E., Hough, B. R., Britten, R. J. and Davidson, E. H. (1976) "Evolution of repetitive and non-repetitive DNA" Chap. 12 of Molecular Evolution (F. J. Ayala, ed.) Sunderland, MA: Sinauer Assoc. pp 200-224

Goldberg, R. B., Crain, W. R., Ruderman, J. V., More, G. P., Barnett, J. R., Higgins, R. C., Gelfand, R. A., Galau, G. A., Britten, R. J. and Davidson, E. H. (1975) "DNA sequence organization in the genomes of five marine invertebrates" Chromosoma (Berl.) 51:225-251

Graham, D. E., Neufeld, B. R., Davidson, E. H. and Britten, R. J. (1974) "Interspersion of repetitive and non-repetitive DNA sequences in the sea urchin genome" Cell 1:127-137

Holmes, D. S. and Bonner, J. (1974) "Sequence composition of rat nuclear deoxyribonucleic acid and high molecular weight ribonucleic acid" Biochemistry 13:841-848

Noll, H. (1967) "Characterization of molecules by constant velocity sedimentation" Nature 215:360-363

Pearson, W. R., Davidson, E. H. and Britten, R. J. (1977a) "A program for least squares analysis of reassociation and hybridization data" submitted to Nuc. Acids Res.

Pearson, W. R., Wu, J. R. and Bonner, J. (1977b) "Analysis of rat repetitive DNA sequences" Thesis California Institute of Technology

Schmid, C. W. and Deininger, P. L. (1975) "Sequence organization of the human genome" Cell 6:345-358

Sober, H. A. (1968) "Deoxyribonucleic acid content per cell of various organisms: Table X Mammals" Handbook of Biochemistry Cleveland: Chemical Rubber Co. p. H-58

Studier, F. W. (1965) "Sedimentation studies on the size and shape of DNA" J. Mol. Biol. 11:373-390

CHAPTER III

ABSENCE OF SHORT PERIOD INTERSPERSION OF REPETITIVE AND NON-REPETITIVE SEQUENCES IN THE DNA OF *DROSOPHILA MELANOGASTER*

Chromosoma (Berl.) 56, 309–326 (1976)

CHROMOSOMA

Absence of Short Period Interspersion of Repetitive and Non-repetitive Sequences in the DNA of *Drosophila melanogaster*

William R. Crain[1], Francine C. Eden[1,2], William R. Pearson[1], Eric H. Davidson[1] and Roy J. Britten[1]*

[1] Division of Biology, California Institute of Technology, Pasadena, California 91125;
[2] present address: Department of Zoology, Indiana University, Blovmington, Indiana 47401, U.S.A.

Abstract. A sensitive search has been made in *Drosophila melanogaster* DNA for short repetitive sequences interspersed with single copy sequences. Five kinds of measurements all yield the conclusion that there are few short repetitive sequences in this genome: 1) Comparison of the kinetics of reassociation of short (360 nucleotide) and long (1,830 nucleotide) fragments of DNA; 2) reassociation kinetics of long fragments (2,200 nucleotide) with an excess of short (390 short nucleotide) fragments; 3) measurement of the size of S1 nuclease resistant reassociated repeated sequences; 4) measurement of the hyperchromicity of reassociated repetitive fragments as a function of length; 5) direct assay by kinetics of reassociation of the amount of single copy sequence present on 1,200 nucleotide long fragments which also contain repetitive sequences.

Introduction

Interspersion of short repetitive sequences with single copy sequences of one to a few thousand nucleotides appears to be a very widespread—even nearly universal—feature of higher animal genomes. DNAs of fifteen organisms widely distributed on the phylogenetic tree have now been examined. In the typical case a majority of the genome consists of short (300 nucleotide pairs) repeated sequences adjacent to longer (1,000–2,000 nucleotide pairs) single copy sequence (recently reviewed by Davidson et al., 1975). *Drosophila melanogaster* DNA does not share this pattern of sequence organization (Manning et al., 1975) and at this time two other exceptions have been found among the insects (Crain, in preparation; Manning, personal communication). These differences are of conceptual importance since DNA sequence organization probably has significant implications for the functional organization of the genome and for genetic regulation (Britten and Davidson, 1969, 1971; Davidson and Britten, 1973).

* Also Staff Member, Carnegie Institution of Washington, Washington, D.C. 20015, U.S.A.

Previous studies of the *Drosophila* genome (Manning et al., 1975) were done principally by electron microscopy of reassociated DNA, while most of the data obtained on more typical genomes have been derived from other methods such as hydroxyapatite assay of reassociation of short and long fragments and S1 nuclease resistance. The question obviously arises as to whether the differences between *Drosophila* DNA and other DNAs are more apparent than real; that is, due to differences in experimental approach. Studies of sequence arrangement in the *Xenopus* genome by electron microscopy (Chamberlin et al., 1975) have yielded results which are entirely consistent with those obtained by other methods. *Xenopus* DNA has a typical pattern of short period sequence interspersion and is the type organism, since its DNA was the first to be studied in detail (Davidson et al., 1973). An additional effort to determine sequence organization in the *Drosophila* genome by several methods other than electron microscopy appeared justified to us. Not only is *Drosophila* an important experimental species, but conceivably the application of other methods might uncover short repeated sequences missed in the electron microscopic measurements. In addition earlier measurements on *Drosophila* DNA (Wu et al., 1972) had been differently interpreted. In this paper we report a search for short repeated sequences in the *Drosophila* genome by a combination of five different methods.

Manning et al. (1975) concluded that the number average length of the repeated sequences in *Drosophila* DNA is about 5,500 nucleotides and that these are probably terminated by single copy sequences. They estimated that there are 2,800 repeats in the genome. From these results we may estimate the quantity of single copy sequence expected to be linked to repeated sequences by a calculation described in the Discussion. If there were a significant number of previously unobserved short interspersed repetitive sequences the measurements reported here would show a larger fraction of single copy DNA linked to repetitive sequences than suggested by Manning et al. The conclusions from our measurements are in substantial agreement with those of Manning et al. (1975).

Material and Methods

1. DNA Preparation. Unlabeled DNA was prepared from syncytium stage eggs that had been stored at −70° C. The eggs were dechorionated by suspension in 50% Chlorox for 2 min at room temperature (2–3 ml/g tissue). They were then caught on a Nitex screen and washed sequentially with distilled H_2O, 70% EtOH and H_2O. The eggs were gently removed from the screen and allowed to settle three times in ice-cold saline-Triton (0.9%NaCl, 0.1% Triton-100; 10 ml/g tissue). The dechorionated eggs (Elgin and Hood, 1973) sink in this mix. The eggs that did not sink were washed again with distilled H_2O. The saline-Triton washed eggs were then washed once more with H_2O (10 ml/g tissue) and suspended in 0.05 M Tris-maleate buffer, pH 7.4; 0.005 M $MgCl_2$ (5 ml/g tissue). The eggs were homogenized with two strokes in a Dounce homogenizer and the homogenate filtered through two layers of Miracloth. The material that was trapped on the Miracloth was resuspended in a small volume of Tris-maleate buffer, rehomogenized in the Dounce homogenizer and filtered through Miracloth again. The filtered material was centrifuged at 2,000 rpm for 10 min through a $^1/_3$ volume sucrose cushion (0.2 M sucrose in Tris-maleate buffer) to pellet the nuclei. The pellet was resuspended in one-half the original volume of Tris-maleate buffer plus 0.1% Triton, and centrifuged a second time through a sucrose cushion. The nuclear pellet was resuspended in 0.1 M Tris, 0.1 M EDTA, pH 8.0 (1 ml buffer to 1 g starting material) and the nuclei lysed in 1% SDS.

The sample was treated with 100 μg/ml predigested pronase (B grade, Calbiochem) for 2 h at room temperature. The mixture was brought to 1.0 M $NaClO_4$ and extracted with an equal volume of phenol: IAC (1:1) and IAC (24:1 chloroform:isoamyl alcohol). The DNA was spooled after precipitation with two volumes of 95% EtOH and redissolved in 0.015 M NaCl, 0.05 M Tris, 0.01 M EDTA, pH 7.6. The DNA was further purified by treatment with 50 μg/ml RNase A (chromatographically pure, Worthington Biochemical Co.) for 1 h at 37° C. The solution was adjusted to 0.1 M NaCl and incubated with 200 μg/ml pre-digested pronase at 37° C for 1.5 h. After digestion the mixture was extracted with an equal volume of IAC and reprecipitated with ethanol. The yield of DNA was approximately 300 μg/g eggs. As an additional purification step sheared DNA was bound to hydroxyapatite at 60° C and washed with 0.12 M phosphate buffer followed by elution with 0.4 M phosphate buffer.

Tritium labeled *Drosophila* DNA purified from tissue culture cells was generously provided to us by Drs. Ray White and David Hogness. A 60 ml suspension of cells that had been grown in the presence of ^{3}H-thymidine was centrifuged and the cells washed in medium without serum. The pellet was suspended in 0.5 ml medium without serum and brought to 10.5 ml with 0.5 M EDTA, pH 9.55; 2.5% sarcosyl; followed by incubation at 60° C. After 2 h 11 ml of 0.1 M EDTA containing 1 mg/ml proteinase K (EM Laboratories) was added and the mixture was incubated at 37° C for 7 h. CsCl was added (1.21 g/ml) and the samples were centrifuged at 33,000 rpm for 3 days at 15° C in a Spinco 40 rotor. The fractions containing the tritium counts were pooled and dialyzed extensively against 0.1 M NaAcetate. The ^{3}H label was shown to be greater than 97% sensitive to DNase and completely resistant to RNase. All sheared tracer preparations were subjected to one further purification step by binding to and elution from hydroxyapatite. The purified tracer melted with an optical hyperchromicity of 95% of that expected for pure DNA. Its specific activity was determined to be 1×10^5 cpm/μg.

2. Preparation of DNA Fragments. DNA fragments of desired lengths were obtained by shearing in a Virtis 60K homogenizer as previously described (Britten et al., 1974).

3. Sizing of DNA Fragments. The single strand length of the DNA fragments was determined on isokinetic alkaline sucrose gradients as described by Noll (1967). The parameters were as follows: $V_{mix}=9.84$ ml, $C_{res}=43\%$ w/v, $C_{flask}=16\%$ w/v in 0.1 M NaOH. The gradients were centrifuged at 41,000 rpm for 20–24 h at 20° C in the Spinco SW41 rotor. The weight average fragment length was determined according to the equations of Studier (1965) relative to markers sized by electron microscopy.

4. DNA Reassociation. Unless otherwise indicated the DNA was reassociated in 0.12 M phosphate buffer at 60° C or 0.4 M phosphate buffer at 66° C. C_0t values determined in the higher salt were corrected to "equivalent" C_0t by multiplying by 4.9. All buffers and DNA solutions were passed through BioRad Chelex 100 equilibrated in the desired buffer. The fraction of DNA molecules containing double strand regions was assayed by hydroxyapatite chromatography (Britten et al., 1974).

The reassociation of DNA was monitored in some experiments by measuring the hypochromicity of denatured DNA samples at 260 nm in 0.12 M phosphate buffer in a water jacketed cuvette in a Beckman ACTA Mark III spectrophotometer.

5. S1 Nuclease Digestion and Sizing of Resistant Regions. Reassociated DNA was adjusted to 0.15 M NaCl, 0.025 M NaAcetate, 0.005 M PIPES, 0.1 mM $ZnSO_4$, 0.005 M β-mercaptoethanol, pH 4.4, and digested for 45 min at 37° C with enough single strand specific S1 nuclease to digest all single strand regions (Ando, 1966). The reaction was terminated by addition of phosphate buffer to a final concentration of 0.12 M and S1 resistant duplex regions were collected on hydroxyapatite at 60° C. ^{14}C labeled sea urchin DNA incubated to appropriate C_0t was included as an internal control in each S1 digestion experiment. After elution from hydroxyapatite with 0.4 M phosphate buffer the resistant regions were chromatographed on a BioRad Biogel A-50 column (1.5 × 110 cm) in 0.12 M phosphate buffer. The column was poured around a support of 6 mm glass beads. An exclusion marker of long native DNA and an inclusion marker of $^{32}PO_3$ were used to calibrate the column.

Results

1. Reassociation Kinetics of Drosophila DNA

Figure 1 shows the reassociation kinetics of two different fragment sizes of *Drosophila melanogaster* DNA assayed by binding to hydroxyapatite. The curves show that *Drosophila melanogaster* has a relatively small quantity of middle

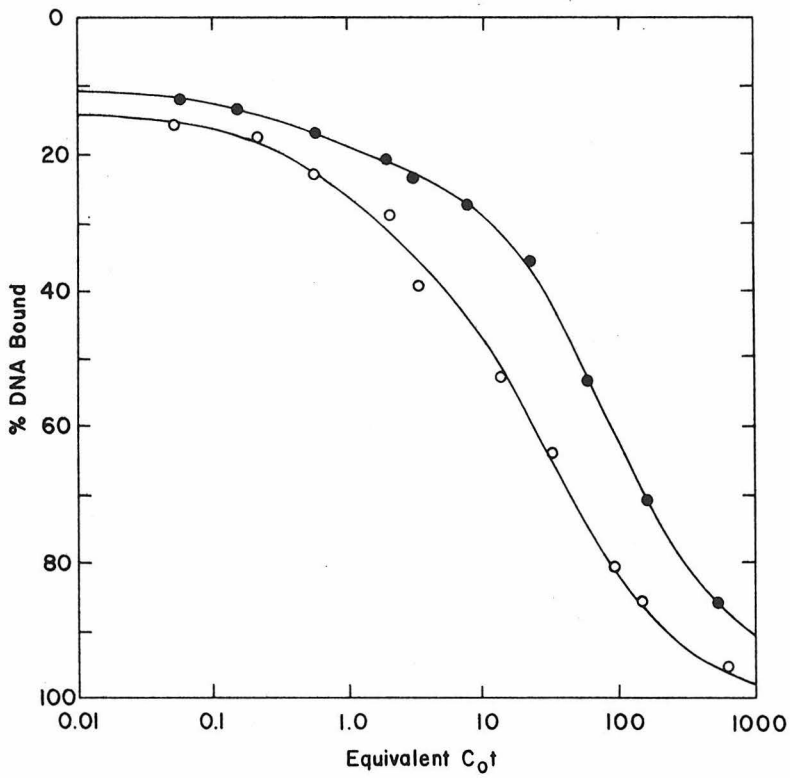


Fig. 1. Reassociation kinetics of *Drosophila* DNA. *Drosophila* DNA fragments of 360 (●) and 1,830 (○) nucleotides were prepared by mixing unlabeled DNA from syncytium stage eggs with ^{3}H-labeled DNA from *Drosophila* tissue culture cells (mass ratios of 1,500/l and 750/l respectively). These mixtures were sheared in the Virtis homogenizer (Britten et al., 1974). Reassociation and assay of the fraction of fragments containing double stranded regions was carried out by binding to hydroxyapatite (see Materials and Methods).

The least squares solution of the data yielded components with the following fraction of the genome (Q) and rate constants (k): For the 360 nucleotide fragments, very highly repetitive sequences and foldback: $Q=0.12$, k cannot be clearly defined; middle repetitive sequences, $Q=0.108$, $k=2.1\ M^{-1}s^{-1}$; single copy sequences, $Q=0.751$, $k=0.0121\ M^{-1}s^{-1}$.

Least squares analysis on the 1830 nucleotide fragment data was carried out with the rate constant for single copy sequence fixed to a value of $0.0273\ M^{-1}s^{-1}$. This value was arrived at by adjusting the k for single copy sequence from that measured with 360 nucleotide fragments. Thus $k_{1,830}=k_{360}\left(\frac{1,830}{360}\right)^{1/2}=0.0273$. The values for the components in the 1,830 nucleotide curve are as follows: very highly repetitive and foldback $Q=0.15$, k not defined; middle repetitive, $Q=0.22$, $k=0.94\ M^{-1}s^{-1}$; single copy, $Q=0.64$, $k=0.0273\ M^{-1}s^{-1}$

repetitive DNA, as previously observed (Laird, 1971; Wu et al., 1972; Schachat and Hogness, 1973; Manning et al., 1975). The kinetic curves are dominated by the single copy DNA reassociation. About 10% of *Drosophila* DNA binds at very early times to hydroxyapatite. This is probably composed of satellites and inverted repeats (Gall et al., 1971; Peacock et al., 1973; Wilson and Thomas, 1974; Manning et al., 1975; Schmid et al., 1975). We have not investigated this fraction in the present work.

The fraction of the DNA fragments containing middle repetitive sequences or only single copy sequences was determined by least squares analysis. These quantities and the most likely rate constant are shown in the legend to Figure 1. The reassociation of the fragments averaging 1,830 nucleotides is somewhat faster than that of the 360 nucleotide fragments. Such a result is expected for the following reasons. First, for those fragments which are purely repetitive or purely single copy the rate of reassociation is increased approximately in proportion to the square root of the fragment size (Wetmur and Davidson, 1968; Wetmur, 1976). Some of the fragments from interspersed regions contain both repetitive and single copy sequences linked together. These fragments are bound to hydroxyapatite when the repetitive regions are reassociated, and therefore when longer fragments are used a greater fraction of the single copy DNA is bound at a given C_0t. This process supplies one of the major methods for the detection and measurement of short period interspersion of repeated and single copy sequences. The rate of reassociation of long and short fragments has been measured for the DNA of a number of species (Davidson et al., 1973; Goldberg et al., 1975; Davidson et al., 1975; Angerer et al., 1975). In each case in which other methods show the presence of short period interspersion the reassociation rate of fragments a few thousand nucleotides long is observed to be strikingly faster than that of fragments a few hundred nucleotides long. However for *Drosophila melanogaster* DNA the effect of increased fragment length is relatively modest. The rate of reassociation of the 1,830 nucleotide fragments is faster than that of the 3,160 nucleotide fragments by a factor which is only slightly larger than the ratio of the square roots of the fragment lengths. This is true for both the repetitive and single copy parts of the reaction. It follows that only a small fraction of the 1,830 nucleotide long fragments are made up of linked repetitive and single copy sequences. We do not yet know whether the rate of reassociation is exactly proportional to the square root of length for fragments as short as 360 nucleotides, and thus a quantitative examination of such measurements is limited in accuracy. The accuracy of this method is not sufficient to detect a small amount of relatively short interspersed repetitive sequences if they exist.

2. Reassociation of Long Labeled Fragments with an Excess of Short Fragments

The reassociation of long labeled fragments with an excess of short fragments was the first method used to demonstrate repetitive sequence interspersion (Britten and Smith, 1970) and it has been used for the measurements of the length of the interspersed single copy sequences in the DNA of several species (e.g.,

Davidson et al., 1973; Graham et al., 1974; Angerer et al., 1975; Efstradiatis et al., 1976). Figure 2 shows the kinetics of reassociation of labeled 2,200 nucleotide long *Drosophila* DNA fragments with excess 390 nucleotide unlabeled fragments. As above the rate of reassociation of the long fragments is affected by their length, so that a small amount of sequence interspersion is difficult to detect. Previous measurements with bacterial DNA fragments (Davidson et al., 1973) indicated that the rate of reassociation of long labeled fragments with an excess of short fragments is directly proportional to the length of the long fragments. The least squares solutions for the rates and quantities of the repetitive and single copy components are shown in the legend to Figure 2. The single copy rate is in fact accelerated by just the ratio of the lengths (2,200/390). Thus the majority of the single copy sequences are not linked to repetitive sequences on fragments of 2,200 average nucleotide length.

In Figure 1 the quantity of the repetitive component appears somewhat larger for the 2,200 nucleotide labeled fragments than for the 360 nucleotide fragments. This is in part due to linkage of single copy sequences to the ends of long repetitive sequences. However the accuracy of these measurements is not sufficient to determine whether the quantity is greater than that expected from the electron microscopic measurements of Manning et al. (1975) (see Discussion).

It is possible that short, interspersed repetitive sequences exist but are either very short or very divergent and thus cannot be bound to hydroxyapatite at standard criterion. If this were true it might be possible to detect these less stable sequences by lowering the criterion of reassociation. To test this possibility the incubations and hydroxyapatite assays were done in 0.12 M phosphate buffer at 50° C, 10° C lower than the standard criterion. Under these conditions more divergent or shorter repeats would reassociate and bind to hydroxyapatite than at the 60° C criterion used in the preceding experiments. The lower curve of Figure 2 shows the result. Additional binding occurred at very low C_0t but no substantial change is observed in the middle repetitive and single copy regions of the kinetics. The implication of the low C_0t binding is that under these conditions a greater amount of DNA was associated with foldback or satellite sequences. Since in *Drosophila* (Schmid et al., 1975) and other species the foldback or palindrome (Wilson and Thomas, 1974) sequences are interspersed throughout much of the genome it is likely, though not proved here, that we are seeing the effect of a larger class of such interspersed sequences. There is, however, no sign of an increased amount of middle repetitive sequences. Thus no evidence for unusually divergent or short interspersed middle repetitive sequences can be obtained by lowering the criterion 10° C.

3. *The Size of S1 Nuclease Resistant Reassociated Repetitive Sequences*

Under appropriate conditions (Britten, Graham, Eden, Painchaud and Davidson, in preparation) the single strand specific nuclease S1 from *Aspergillus* (Ando, 1966; Vogt, 1973; Sutton, 1971) is useful for determining the length of repetitive sequences. With a carefully controlled digestion single strands can

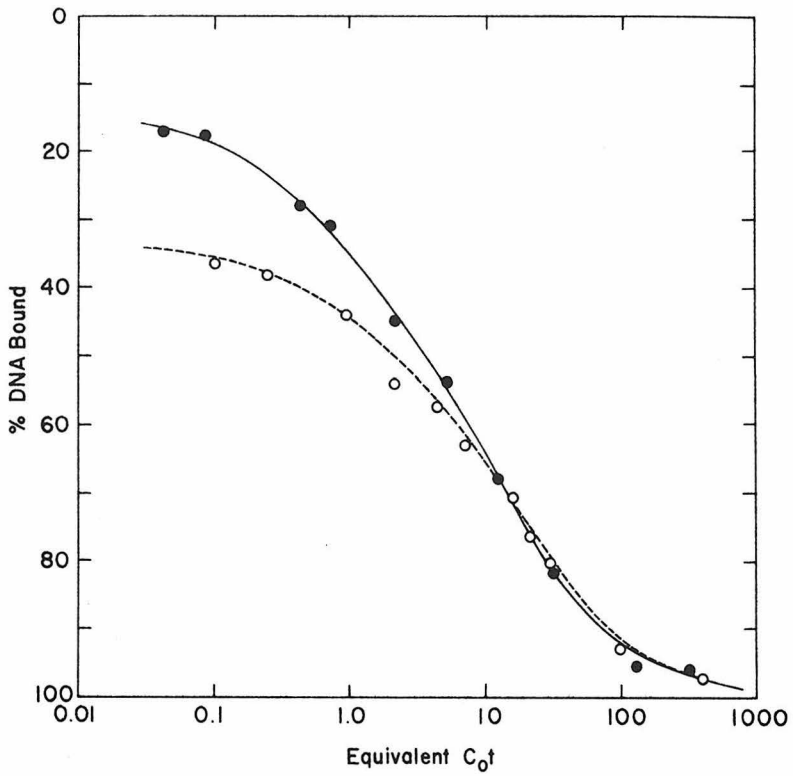


Fig. 2. Reassociation kinetics of 2,200 nucleotide labeled fragments in the presence of an excess of 390 nucleotide DNA fragments. The upper curve (●) shows the binding of the labeled fragments to hydroxyapatite at 60° C in 0.12 M phosphate buffer after incubation under the same conditions. The lower curve (○) shows the binding to hydroxyapatite at 50° C in 0.12 M phosphate buffer after incubation under these conditions.

Least squares analysis was performed on the data at 60° C and 50° C yielding the following components:

60° C. Very highly repetitive and foldback, $Q=0.17$, k not defined; middle repetitive, $Q=0.27$, $k=1.66\ M^{-1}s^{-1}$; single copy, $Q=0.584$, $k=0.074\ M^{-1}s^{-1}$.

50° C. Very highly repetitive and foldback, $Q=0.34$, k not defined; middle repetitive, $Q=0.166$, $k=1.06\ M^{-1}s^{-1}$; single copy, $Q=0.497$, $k=0.048\ M^{-1}s^{-1}$

be effectively removed while the enzyme does not attack duplexes containing even 10% to 20% mismatched base pairs. The following procedure has been developed to measure the length distribution of short repetitive sequences: a) reassociation of approximately 2,000 nucleotide long fragments to a C_0t at which few single copy sequences are reassociated; b) limited digestion with S1 nuclease to cleave single strands from duplexes; c) binding to hydroxyapatite (0.12 M phosphate buffer, 60° C) to separate duplexes from undigested single strands; d) gel filtration analysis of the S1 resistant duplexes on agarose A-50 columns. This method has been applied to the DNA of a number of species which possess typical short interspersed repetitive sequences. In every case the resistant duplexes separate into two size classes: a long fragment class, which is excluded from the agarose column; and a short fragment class with a mode

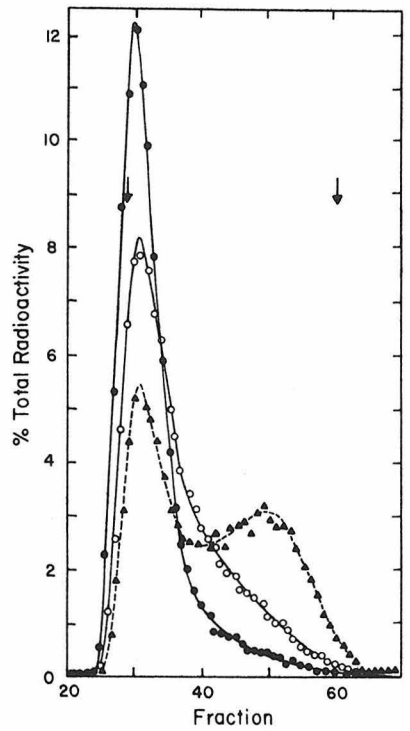


Fig. 3. Length of S1 resistant duplexes formed by renaturing highly repetitive and middle repetitive *Drosophila* DNA. Highly repetitive and middle repetitive DNA sequences were prepared from 2,400 nucleotide fragments as described in text and treated with S1 nuclease sufficient to remove single strands without digesting mismatched regions. 436 μg of highly repetitive and 600 μg of middle repetitive DNA were digested under the following conditions; 0.15 M NaCl, 0.005 M PIPES, 0.025 M Na Acetate, pH 4.3, 0.1 mM $ZnSO_4$, 37° C, 45 min. The resistant duplex fraction was collected on hydroxyapatite and chromatographed on agarose A-50 in 0.12 M phosphate buffer. Sea urchin DNA (2,000 nucleotides) reassociated to C_0t 4 was added to the *Drosophila* DNA before S1 digestion as a control. The arrows indicate the positions of exclusion and inclusion markers. ● *Drosophila* satellite DNA (C_0t 0.05 bound). ○ *Drosophila* middle repetitive DNA (C_0t 0.05–3.0), ▲ Sea urchin repetitive DNA

size of about 300 nucleotide pairs (Goldberg et al., 1975; Britten, Graham, Eden, Painchaud and Davidson, in preparation). The fraction of the repeats which fall in the short repetitive sequence class varies from 35–85%, depending on the species (Davidson et al., 1975).

In order to achieve greater sensitivity the *Drosophila* DNA fragments were partially separated into a very rapidly reassociating fraction and a middle repetitive fraction. For this purpose the 2,400 nucleotide long fragments were incubated to C_0t 0.05 and passed over hydroxyapatite. The bound fraction was eluted with 0.4 M phosphate buffer. The unbound fraction was reincubated to C_0t 3. Both fractions were digested with S1 nuclease and the size of the resistant duplex structure measured by gel filtration on agarose A-50. Details of the procedure and the yield of DNA in each fraction are given in Materials and Methods. The size distributions of the S1 resistant duplexes are shown in Figure 3

along with a control, labeled sea urchin DNA. Identical patterns were obtained for the sea urchin DNA in both runs. The yield of the short fragments of sea urchin DNA is a little less than the amount expected but the S1 digestion was clearly effective. The C_0t 0.05 bound fraction of the *Drosophila* DNA shows no measurable quantity of short S1 resistant duplexes. The middle repetitive fraction (C_0t 0.05 to 3) appears to show a small quantity of short resistant duplexes. About 25% of the DNA has a k_{av} greater than 0.3, or a size less than 1,000 nucleotide pairs. This result is entirely consistent with the pattern of long repetitive sequences measured by the electron micrographic measurements of Manning et al. (1975).

4. The Hyperchromicity of Reassociated Repetitive Fragments

When DNA fragments containing both repetitive and single copy sequences are reassociated to low C_0t the single copy regions remain unpaired. The quantity of these unpaired regions can be measured by collecting all of the fragments containing paired regions on hydroxyapatite and measuring their hyperchromicity.

The amount of single copy DNA linked to repetitive sequences on such fragments depends on the sequence interspersion pattern. If there are a large number of short interspersed repeats a large quantity of single copy sequences

Table 1. Hyperchromicity of middle repetitive DNA sequence after reassociation

Preparation of middle repetitive DNA fraction

Initial average fragment length (nucleotides)	350	1,600	3,200
First incubation			
C_0t	3	3	3
Fraction bound	0.24	0.31	0.32
Fraction to total DNA	0.24	0.31	0.32
Second incubation			
C_0t	0.012	0.016	0.016
Fraction unbound	0.41	0.55	0.56
Fraction of total DNA	0.096	0.17	0.18
Third incubation			
C_0t	0.29	0.51	0.54
Fraction bound	0.44	0.30	0.22
Fraction of total DNA	0.042	0.05	0.039
Hyperchromicity determination			
Fourth incubation (for spectrophotometric melt)			
C_0t	12.6	15.3	11.7
Final fragment size (nucleotides)	300	1,000	1,600
Hyperchromicity	0.22	0.20	0.19
Fraction of native DNA hyperchromicity	0.95	0.83	0.79
T_m (°C)	78.1	83.5	84.5

will be linked to them, and a low hyperchromicity will be observed. The amount of hyperchromicity decreases as the length of the fragments increases. This effect has been demonstrated in measurements of the hyperchromicity of partially reassociated DNA fragments in several species that display typical interspersion of short repetitive sequences (Davidson et al., 1974; Graham et al., 1974; Goldberg et al., 1975; Angerer et al., 1975).

The results for the middle repetitive fraction of *Drosophila* are listed on the lower part of Table 1. The upper parts of this table show the incubations and hydroxyapatite fractionation steps used in purifying the appropriate middle repetitive fraction of *Drosophila* DNA. The reassociated fragments which are 300 nucleotides long at the end of the procedure have about 95% of the hyperchromicity of native DNA, indicating that very little single copy sequence is linked to them. Even the longer fragments (1,000 and 1,660 nucleotides) show hyperchromicities of 83% and 79% of native DNA. These values suggest that some single copy DNA is linked to repeats in fragments of this size and indeed some is expected if the average repetitive sequence is 5,000 nucleotides long (Manning et al., 1975).

Table 2. Estimates from reassociation kinetics of the components of *Drosophila* DNA

Component	Single copy	Middle repetitive	Fast
Data presented in this paper			
Quantity	0.74	0.09	0.14
Repetition frequency	1.0	30	–
Complexity	6.7×10^7 NTP	3.14×10^6	
Rate constant in whole DNA	0.0121[a]	0.35[b]	–
Rate constant if pure	0.0164	3.89	
Data of Manning et al.[c]			
Quantity	0.70	0.12	0.12
Repetition frequency	1	70	–
Rate constant in whole DNA	0.0069	0.5	–

[a] In the fit described here the single copy rate constant was fixed to agree with that of the free fit for the 360 nucleotide fragment reassociation curve

[b] For this fit the rate constant for the middle repetitive component in the 1,830 and 2,200 nucleotide curves were corrected to their predicted value for 360 nucleotide fragments and the average rate constant determined. This average middle repetitive rate constant was then fixed in the least squares fit of the 360 nucleotide fragment reassociation curve. The predicted rate constants for 360 nucleotide fragments from the two longer fragment size curves were calculated as follows: For 1,830 nucleotide fragments, where $0.936\ M^{-1}\ s^{-1}$ is the rate observed for the middle repetitive sequences in the experiments of Fig. 1.

$$k = 0.936 \left(\frac{360}{1{,}830}\right)^{1/2} = 0.45.$$

For 2,200 nucleotide fragments driven by 390 nucleotide fragments, where $1.66\ M^{-1}s^{-1}$ is the rate observed for the middle repetitive sequences in the experiment of Fig. 2.

$$k = 1.66 \left(\frac{390}{2{,}200}\right)\left(\frac{360}{390}\right)^{1/2} = 0.283.$$

The average of these two rate constants is $0.35\ M^{-1}s^{-1}$.

[c] Fits of data from Manning et al. (1975) for *Drosophila* DNA fragments of 400 nucleotides length

5. Kinetic Assay for Single Copy DNA Sequences Linked to Repetitive Sequences

The following is probably the most direct and most sensitive of the methods used in this paper to determine the amount of single copy DNA sequences which are linked to repetitive DNA sequences. Long labeled DNA fragments which contain repetitive sequences were selected. These fragments were sheared and reassociated with an excess of total *Drosophila* DNA fragments. The reassociation kinetics were analyzed for evidence of fragments bearing only single copy sequences. The sensitivity of this experiment is such that a few percent of single copy DNA can be detetced.

The first step was to isolate a large quantity of short repetitive DNA fragments to be used to select the long labeled fragments containing repeats. The frequency with which middle repetitive sequence occur in the *Drosophila* genome is relatively low, about 30 copies of each sequence. This agrees well with the previously reported repetition frequency of the middle repetitive component (Wu et al., 1972; Schachat and Hogness, 1974; Manning et al., 1975). This makes it difficult to separate single copy from repetitive DNA and a single fractionation step does not suffice. Table 3 show the fractionation steps and yield at each stage in comparison with that expected from our best estimates of the repetitive frequency components of *Drosophila* DNA (Table 2). Twenty mg of *Drosophila* DNA were sheared to 300 nucleotides and 14% was recovered

Table 3. Fractionation of *Drosophila* repetitive DNA

C_0t	Fraction bound to hydroxyapatite	Fraction unbound	Predicted fraction[a] bound	Predicted fraction[a] unbound
40	0.47		0.47	
20	0.63		0.54	
0.005		0.724		0.63
5	0.74			0.73

[a] The fraction of the DNA bound to hydroxyapatite at the different steps in the fractionation was arrived at by calculating the theoretical fraction reassociated for each component according to the equation

$$\frac{C}{C_0} = \frac{1}{1 + k C_0 t}$$

where C is the concentration of totally single stranded fragments, C_0 is the total DNA nucleotide concentration, t is time and k is the second order rate constant. The predicted recoveries for the individual components were added to give the predicted total fractions bound or unbound, as shown in columns 4 and 5. The rate constants of each component were adjusted according to the new concentration in the remaining DNA in order to calculate the recovery in the next step. This calculation was made using the components arrived at from these data: 14% zero time binding and highly repetitive DNA; 9% middle repetitive sequences with a rate constant of 0.35 M^{-1} s^{-1}, and 74% single copy with a rate constant at 0.0121 M^{-1} s^{-1}. The rate constants were adjusted to their values for 290 nucleotide length fragments and a rate constant of 100 M^{-1} s^{-1} was estimated for the fast component

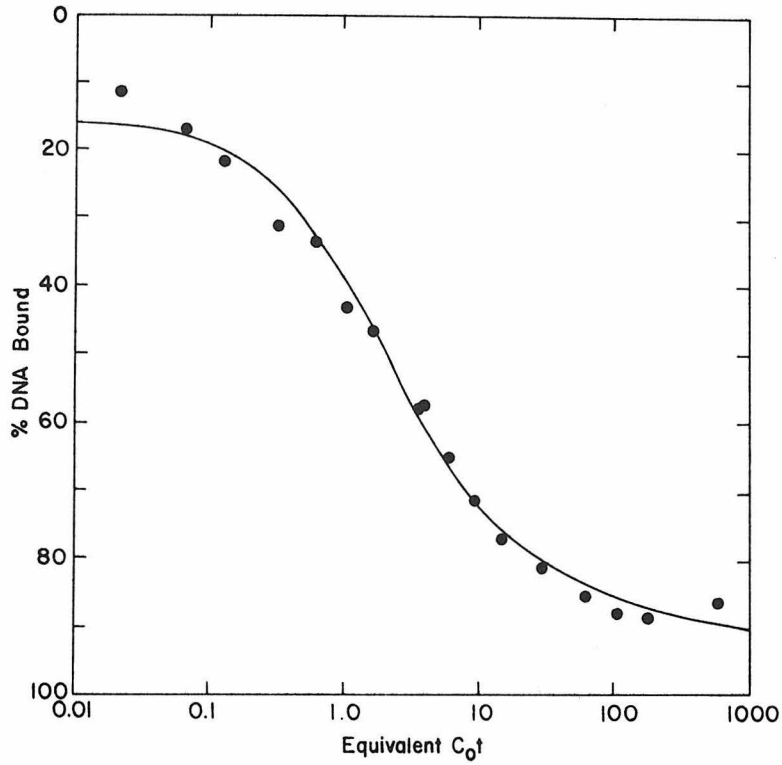


Fig. 4. Reassociation of sheared selected repetitive DNA in the presence of excess whole DNA. Tritium labeled *Drosophila* DNA (tracer) fragments which had been stripped of zero time binding sequences were reassociated to C_0t 3.1 in the presence of a 50-fold excess of 290 nucleotide purified *Drosophila* repetitive DNA (driver). The tracer which had reassociated with the repetitive driver was collected by hydroxyapatite chromatography. The single strand size of the tracer after elution from hydroxyapatite was 1,225 nucleotides. The DNA that bound to hydroxyapatite was recovered and sheared, reducing the tracer length to 375 nucleotides. The sheared material was mixed with 290 nucleotide whole *Drosophila* DNA in a ratio of 1 to 5 and reassociated to various C_0t values. The fraction of fragments containing duplexed regions was measured on hydroxyapatite. The curve represents a least squares fit to the data with the rate constant for the single copy component fixed at 0.012 $M^{-1}s^{-1}$ according to the following description.

The rate of the single copy component for the 375 nucleotide fragment tracer in this curve was arrived at by adjusting the rate for 360 nucleotide fragments as follows:

$$k_{290}=k_{360}\left(\frac{290}{360}\right)^{1/2}\left(\frac{375}{290}\right)=0.014\ M^{-1}s^{-1}.$$

This rate constant was further adjusted to allow for the dilution of the single copy sequences when whole DNA driver was mixed with purified repeat driver. Assuming that the repeat driver contained no single copy sequences the single copy sequences would be diluted by a factor of 0.833. The resulting rate constant for single copy sequences is thus:

$$0.833\ (0.014)=0.012\ M^{-1}s^{-1}$$

as the repetitive fraction. This should consist of 34% highly repetitive, 66% middle repetitive and 0.2% single copy DNA, were there no sequence interspersion in this genome. The reassociation kinetics of this fraction were measured in the spectrophotometer by optical hyperchromicity. The results were consistent

with the expected components and the reassociation went to completion (final optical density = 0.75 of the initial denatured A_{260}). No single copy component could be observed in the reaction kinetics.

The second step was to prepare long labeled fragments. ^3H labeled DNA was sheared to 1,800 nucleotide pairs, denatured, incubated to C_0t 0.005 and passed through hydroxyapatite to remove the zero time binding or foldback fraction. This step was repeated, and a total of 10.2% of the fragments were removed. The remaining fragments were incubated with a 50 fold excess of the short repetitive fragments (described above) to C_0t 3.2. The sample was then passed over hydroxyapatite in 0.12 M phosphate buffer at 60° C. Twenty-seven percent of the long labeled fragments and 79% of the short driver fragments were bound. The bound fragments were eluted at 98° C and the tracer recovered was 1,225 nucleotides long, as assayed by alkaline sucrose sedimentation. These fragments were sheared to 375 nucleotides and incubated with a 5 fold excess of total *Drosophila* DNA. The kinetics of the reassociation are shown in Figure 4. It is clear that the principal portion of the labeled fragments is made up of repetitive DNA sequences. Least squares analyses were carried out to determine the best estimates of the fraction of the sheared tracer fragments which contained only single copy sequence. For this purpose the rate constant for the single copy component was fixed at 0.012 $M^{-1}s^{-1}$, the rate constant of single copy sequence in the driver DNA. All of the other parameters were allowed to remain free. The solution indicates that 9% of the labeled fragments are single copy sequences. We have also made a least squares analysis with the middle repetitive as well as the single copy rate constants fixed at the values shown in Table 2 (corrected for the additional repetitive sequences present in this mixture). With these restrictions the calculated fraction of fragments bearing only single copy sequence is 14%.

Discussion

Our purpose here is to calculate the expected amount of single copy sequence linked to repetitive sequences and to examine the quantitative conclusions from the five types of measurements. In this calculation G is the genome size in nucleotide pairs and L is the fragment length. For a particular set of repetitive sequences (indicated by the subscript 1) S_1 is the length of the interspersed single copy sequences, Q_1 the fraction of the genome in these single copy sequences and F_1 the number of interspersed repetitive or single copy sequences in this set. Figure 5 shows for an ideal case the increase with fragment length of the fraction of fragments containing single copy DNA linked to repeated sequences. The slope (for $L < S_1$) in Figure 5 is $dR/dL = Q_1/S_1$. Since $Q_1 = F_1 S_1/G$ the slope can be written:

$$dR/dL = F_1/G. \tag{1}$$

This simple relationship is for a particular set of interspersed sequences and we may add up the contribution to the slope of each such set as long as the fragment length is less than the length of the single copy sequences.

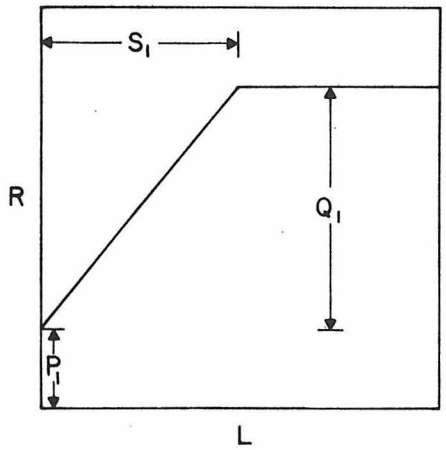


Fig. 5. The contribution of one element of a sequence interspersion pattern to hydroxyapatite binding. Shown as a function of fragment length (L) is R, the amount of DNA in fragments which contain one set of repeated sequences spaced by single copy sequences of length S_1. The total quantity of single copy DNA interspersed with this set is Q_1 and the quantity of repeated DNA is P_1. For short fragments little single copy DNA is linked to the repeats and in the limit only the repetitive DNA (P_1) would be included. The amount of single copy DNA increases in proportion to the fragment length until the length of the single copy sequences is reached (plus enough length to recognize the succeeding repeat). The slope of the linear rising portion is S_1/Q_1

Thus for $L < S$ the total initial slope of the curve for the fraction of fragments containing single copy DNA linked to repetitive sequence is:

$$(dR/dL)_{\text{initial}} = (F_1 + F_2 + F_3 + \ldots F_n)/G = N/G \quad (2)$$

where F_1, F_2 etc. are the numbers of single copy (or repetitive) sequence elements in each interspersed set and N is the total number of interspersed sequences in the genome. For L less than S the fraction of the genome which is in single copy sequences linked to repetitive sequences (Y) follows directly:

$$Y = L(dR/dL) = L\,N/G. \quad (3)$$

Manning et al. (1975) take the genome size of *Drosophila melanogaster* to be 1.3×10^8 nucleotide pairs (Fristrom and Yund, 1973) and estimate that there are 2,800 interspersed repeated sequences with an average length of 5,500 nucleotides. If we assume that all of these are interspersed with single copy sequences that are more than few thousand nucleotides long, then for a 1,000 nucleotide fragment length:

$$Y = 1{,}000 \times 2{,}800/1.3 \times 10^8 = 0.0215.$$

Therefore we except to find about 2% of the genome (2.8×10^6 nucleotide pairs) as single copy sequence linked to fragments 1,000 nucleotides long which contain middle repetitive sequences. From Table 2 the amount of middle repetitive DNA is 9% of the genome or 11.7×10^6 nucleotide pairs. Therefore 1,000 nucleotide long fragments of middle repetitive DNA would contain 19% (2.8×10^6/

Table 4. Fraction of middle repetitive and single copy sequences in *Drosophila* genome

Fragment size	Middle repetitive component			Single copy component		
	Fraction of fragments observed[a]	Fraction of genome that is single copy attached to repeats[b]	Corrected fraction of genome in middle repetitive sequence[c]	Fraction of fragments[a]	Fraction of genome that is single copy attached to repeats[b]	Corrected fraction of genome in single copy sequence[c]
360 nucleotides	0.091	0.008	0.083	0.739	0.008	0.747
1,830 nucleotides	0.221	0.039	0.182	0.637	0.039	0.676
2,200 nucleotides	0.267	0.047	0.220	0.584	0.047	0.631

[a] Quantities of components determined from least squares fit of reassociation data at given fragment length (from Figs. 1 and 2 and Table 2)

[b] Fraction of genome that is single copy sequences attached to repeated sequences at the various fragment lengths. Calculation as described in text

[c] Fraction of genome composed of component after correction for fraction of DNA which is single copy linked to repeats at the fragments length used

$2.8 \times 10^6 + 11.7 \times 10^6$) single copy DNA. We take this to be the best estimate from the electron micrographic measurements of Manning et al. For calculations involving other fragment lengths the quantity of linked single copy DNA is simply proportional to the fragment length as long as the length of typical single copy sequence is not exceeded.

Calculations from the Kinetics of Reassociation (Parts 1 and 2)

Table 4 shows the expected fraction of the single copy DNA calculated as above from the number of middle repetitive sequences (2,800) estimated by Manning et al. (1975). We have subtracted this estimate from the least squares estimates of the fraction of DNA in the middle repetitive kinetic components. If the number of interspersed middle repetitive sequences were significantly greater than 2,800 we would observe, after this subtraction, an increase in the middle repetitive fraction with fragment length. Though small, such an effect is seen in Table 4 and it is this uncertain result which led us to attempt corroboration with other, more sensitive methods.

Measurement of the Size of the S1 Nuclease Resistant Fragments

The size distribution of S1 resistant middle repetitive duplexes is shown in Figure 3 as open circles. The great majority of the fragments are excluded from agarose and are therefore greater than 2,000 nucleotides in length. A small fraction, perhaps 15% of the mass of DNA has a k_{av} greater than about 0.4, and thus their length is less than 400 or 500 nucleotide pairs. This represents

the best estimate by this method of the fraction of the middle repeats of about 300 nucleotide average length. While this is a simple and direct method, the following calculation shows that it is not a sensitive way to detect a small number of short repetitive sequences. If there were, for example, as many 300 nucleotide repetitive duplexes as all others combined we would expect 300/5,600= 5% of the total DNA in this region of the A-50 chromatogram. We obviously can not rule out such a possibility from this measurement. In fact the data suggest a potentially larger number of short repeats. However there are a variety of artifactual possibilities which could give rise to short fragments, for example shear breakage, and internal nicking by the S1 nuclease. We prefer to draw no conclusion from the tail of short fragments shown in Figure 3. In any case it is clear from this measurement that the bulk of the middle repetitive sequences are long.

Hyperchromicity of Reassociated Repetitive Fragments

The results of fractionating *Drosophila* DNA fragments of various sizes are shown on Table 1 along with the hyperchromicity of the purified middle repetitive fractions. The fractionation indicates that only a small increase in single copy DNA linked to middle repetitive DNA occurs with fragment size. The three final fragment lengths are 300, 1,000 and 1,600 nucleotides. The hyperchromicity data imply for these three fragment lengths respectively that 5%, 16.5% and 21% were made up of linked single copy DNA. The calculation described above, assuming that there are just 2,800 interspersed middle repetitive sequences in the genome yields 5%, 14% and 21% for the three lengths. This agreement is surprisingly good considering the possible sources of error in these measurements.

Direct Kinetic Assay for Single Copy Sequences Linked to Middle Repeats

The least squares solutions of the data shown in Figure 4 indicate that 9–14% of the selected 1,225 nucleotide long fragments is made up of linked single copy DNA sequences. The calculation described above yields 17% for this fragment length. This agreement is certainly within error. This measurement is the only one of all of the approaches we have applied to *Drosophila* DNA which directly implies that single copy DNA is interspersed with the repeats at any length. The other measurements have relied, for example, on the increase with length of the fraction of the genome associated with middle repeats or on the fact that the linked DNA was single stranded. In every case it was reasonable but unproved assumption that the linked DNA is actually single copy sequence. The electron micrographic measurements of Manning et al. are subject to the same proviso. However the least squares solution of the data of Figure 4 includes a component with the kinetics of single copy DNA which is of about the quantity expected if there are 2,800 repetitive sequences interspersed with single copy sequences.

Some of the fragments of *Drosophila* DNA that have been "cloned" and examined by Wensink et al. (1974) contain both single copy and repetitive sequences. In addition all of the cloned middle repetitive sequences so far examined are several thousand nucleotides or more in length. The data shown in Figure 4 give the best direct estimate of the number of places in the *Drosophila* genome at which repeated sequences are linked to single copy sequences. The actual number of such transitions in the DNA sequence is, of course, just twice the number of interspersed repetitive and single copy sequences.

Conclusion

Each of the measurements described in this paper gives a result that is consistent, within error, with the conclusion that about 2,800 relatively long middle repetitive sequences are interspersed with single copy sequences in the *Drosophila* genome. We may ask how many more interspersed repetitive sequences could be added to the 2,800 used in the calculations above without creating an unacceptable deviation from the observed measurements.

It seems very unlikely that a doubling of the number would be acceptable. While this would not introduce a problem of interpretation for the kinetics and S1 resistance methods, serious deviations would be revealed by the last two methods described. The hyperchromic shift in one set of measurements and the amount of single copy DNA in the other would have to be in error by a factor of two. We do not believe this is at all likely and conclude that a two fold greater number of interspersed repeats is ruled out. Of course smaller deviations are possible.

Acknowledgements. We wish to thank Dr. Ray White for his generous gift of tritium labeled *Drosophila* DNA. This work was supported by grants from the National Institutes of Health (HD-05753 and GM-20927) and from the National Science Foundation (BMS75-07359). W.R.C. holds a postdoctoral fellowship from the National Institutes of Health (GM-55726). F.C.E. held a postdoctoral fellowship from the American Cancer Society (PF-955). W.R.P. is a predoctoral fellow on a National Institutes of Health training grant (GM-00086).

References

Ando, T.: A nuclease specific for heat-denatured DNA isolated from a product of Aspergillus oryzae. Biochim. biophys. Acta (Amst.) **114**, 158–168 (1966)

Angerer, R.C., Davidson, E.H., Britten, R.J.: DNA sequence organization in the mollusc Aplysia californica. Cell **6**, 29–39 (1975)

Britten, R.J., Davidson, E.H.: Gene regulation for higher cells: a theory. Science **165**, 349–357 (1969)

Britten, R.J., Davidson, E.H.: Repetitive and nonrepetitive DNA sequences and a speculation on the origins of evolutionary novelty. Quart. Rev. Biol. **46**, 111–138 (1971)

Britten, R.J., Graham, D.E., Neufeld, B.R.: Analysis of repeating DNA sequences by reassociation. In: Methods in enzymology (L. Grossman and K. Moldave, eds.), Vol. 29, Part E, pp. 363–418. New York: Academic Press 1974

Britten, R.J., Smith, J.: A bovine genome. Carnegie Inst. Wash. Yearb. **68**, 378–386 (1970)

Chamberlin, M.E., Britten, R.J., Davidson, E.H.: Sequence organization in Xenopus DNA studied by the electron microscope. J. molec. Biol. **96**, 317–333 (1975)

Davidson, E.H., Britten, R.J.: Organization, transcription and regulation in the animal genome. Quart. Rev. Biol. **48**, 565–613 (1973)

Davidson, E.H., Galau, G.A., Angerer, R.C., Britten, R.J.: Comparative aspects of DNA sequence organization in metazoa. Chromosoma (Berl.) **51**, 253–259 (1975)

Davidson, E.H., Graham, D.E., Neufeld, B.R., Chamberlin, M.E., Amenson, C.S., Hough, B.R., Britten, R.J.: Arrangement and characterization of repetitive sequence elements in animal DNAs. Cold Spr. Harb. Symp. quant. Biol. **38**, 295–301 (1974)

Davidson, E.H., Hough, B.R., Amenson, C.S., Britten, R.J.: General interspersion of repetitive with non-repetitive sequence elements in the DNA of Xenopus. J. molec. Biol. **77**, 1–23 (1973)

Efstradiatis, A., Crain, W.R., Britten, R.J., Davidson, E.H., Kafatos, F.: DNA sequence organization in the lepidopteran Antherea pernyi. Proc. nat. Acad. Sci. (Wash.) (in press, 1976)

Elgin, S.C.R., Hood, L.E.: Chromosomal proteins of Drosophila embryos. Biochemistry **12**, 4984–4991 (1973)

Fristrom, J.W., Yund, M.A.: Genetic programming for development in Drosophila. Crit. Rev. Biochem. **1**, 537–570 (1973)

Gall, J.G., Cohen, E.H., Polan, M.L.: Repetitive DNA sequences in Drosophila. Chromosoma (Berl.) **33**, 319–344 (1971)

Goldberg, R.B., Crain, W.R., Ruderman, J.V., Moore, G.P., Barnett, T.R., Higgins, R.C., Gelfand, R.A., Galau, G.A., Britten, R.J., Davidson, E.H.: DNA sequence organization in the genomes of five marine invertebrates. Chromosoma (Berl.) **51**, 225–251 (1975)

Graham, D.E., Neufeld, B.R., Davidson, E.H., Britten, R.J.: Interspersion of repetitive and non-repetitive DNA sequences in the sea urchin genome. Cell **1**, 127–137 (1974)

Laird, C.D.: Chromatid structure: relationship between DNA content and nucleotide sequence diversity. Chromosoma (Berl.) **32**, 378–406 (1971)

Manning, J.E., Schmid, C.W., Davidson, N.: Interspersion of repetitive and non-repetitive DNA sequences in the Drosophila melanogaster genome. Cell **4**, 141–155 (1975)

Noll, H.: Characterization of macromolecules by constant velocity sedimentation. Nature (Lond.) **215**, 360–363 (1967)

Peacock, W.J., Brutlag, D., Goldring, E., Appels, R., Hinton, C.W., Lindsley, D.L.: The organization of highly repeated DNA sequences in Drosophila melanogaster chromosomes. Cold Spr. Harb. Symp. quant. Biol. **38**, 405–416 (1973)

Schachat, F.H., Hogness, D.S.: Repetitive sequences in isolated Thomas circles from Drosophila melanogaster. Cold Spr. Harb. Symp. quant. Biol. **38**m 371–381 (1974)

Schmid, C.W., Manning, J.E., Davidson, N.: Inverted repeat sequences in the Drosophila genome. Cell **5**, 159–172 (1975)

Studier, F.W.: Sedimentation studies of the size and shape of DNA. J. molec. Biol. **11**, 373–390 (1965)

Sutton, W.D.: A crude nuclease preparation suitable for use in DNA reassociation experiments. Biochim. biophys. Acta (Amst.) **240**, 522–531 (1971)

Vogt, V.M.: Purification and further properties of single-strand-specific nuclease from Aspergillus oryzae. Europ. J. Biochem. **33**, 192–200 (1973)

Wensink, P.C., Finnegan, D.J., Donelson, J.E., Hogness, D.S.: A system for mapping DNA sequences in the chromosomes of Drosophila melanogaster. Cell **3**, 315–325 (1974)

Wetmur, J.G.: Hybridization and renaturation kinetics of nucleic acids. Ann. Rev. biophys. Bioeng. **5** (in press, 1976)

Wetmur, J.G., Davidson, N.: Kinetics of renaturation of DNA. J. molec. Biol. **31**, 349–370 (1968)

Wilson, D.A., Thomas, C.A., Jr.: Palindromes in chromosomes. J. molec. Biol. **84**, 115–144 (1974)

Wu, J.-R., Hurn, J., Bonner, J.: Size and distribution of the repetitive segments of the Drosophila genome. J. molec. Biol. **64**, 211–219 (1972)

Received March 2, 1976 / Accepted March 22, 1976 by J.G. Gall
Ready for press March 30, 1976

CHAPTER IV

KINETIC DETERMINATION OF THE GENOME SIZE OF THE PEA

Introduction

In the past few years, substantial data on the organization of repeated and single copy sequences in animal DNAs have been accumulated. There have been relatively few studies on DNA sequence organization in plants (Walbot and Dure, 1976; Smith et. al., 1976). There is evidence that plant DNAs may contain large fractions of repetitive sequences. Studies on the cotton genome (Walbot and Dure, 1976) suggest that the organization of the large repetitive fraction may be different from the typical 300 nucleotide short repeat/1000-2000 nucleotide long single copy spacer pattern seen in most animals. (for review, see Davidson et. al., 1975). Walbot and Dure (1975) have reported that 40% of the cotton genome is repeated and that the repeated sequences average 1250 nucleotides in length while the interspersed single copy sequences are 2000 nucleotides long.

We have measured the fraction of repetitive and single copy DNA in the pea genome. Our measurements on the single copy rate of reassociation suggest that the size of the pea genome is substantially smaller than previous reports (McLeisch and Sunderland, 1961; Van't Hoff and Sparrow, 1963; Birnsteil et. al., 1964). More than 70% of the DNA sequences in the pea are repeated about 500 fold and the rate constant for the single copy reassociation is .00215, corresponding to a genome size of .459 pg. according to our best estimates.

Materials and methods

Isolation of crude chromatin from pea embryos

100 lb of Alaskan peas (Pisum sativum) were germinated for 3 days. 23 lb of pea embryos were isolated using a system for large scale preparation of pea embryo tissue. Embryos were stored at -20 °C.

For preparation of crude chromatin, 4 lb of frozen embryos were homogenized in a 1 gal Waring blender in 0.25 M sucrose 50 mM Tris pH 8.0 1 mM $MgCl_2$ and centrifuged at 5000g in a Sorvall GSA rotor for 30 min. The crude nuclear pellet was separated from a starch pellet, resuspended in grinding buffer and centrifuged at 10,000g for 15 min. The nuclear pellet was again separated from the starch pellet, resuspended in 50 mM Tris pH 8 and centrifuged at 10,000g for 20 min twice.

Isolation of DNA

The crude chromatin pellet was suspended in 0.1 M Tris 0.1 M EDTA pH 8.5 and sodium dodecyl sulfate added to 1%. The viscous mixture was brought to 37 °C and incubated with 50 ug/ml Pronase (Calbiochem grade B) until the DNA went into solution. The DNA was extracted twice with a mixture of phenol:chloroform:isoamyl alcohol (25:24:1) and extracted twice with chloroform:isoamyl alcohol (24:1). The DNA was then precipitated at room temperature in 2.5 volumes of 95% ethanol and allowed to dissolve overnight in 0.05 M Tris 0.01 M EDTA pH 8.5.

Once in solution, the DNA was digested with RNase A (pre-incubated 10 min at 95 °C) at 50 ug/ml for 90 min at 37 °C and digested with 200 ug/ml Pronase (pre-incubated 45 min at 37 °C) for 90 min at 37 °C. The DNA was then extracted with chloroform:isoamyl alcohol (24:1) three times and reprecipitated with ethanol. The DNA was redissolved and reprecipitated with ethanol.

The DNA was sheared to 350 nucleotides by three passes through a french pressure cell at 50,000 psi. and passed over Chelex-100 (BioRad) before renaturation.

Preparation of ^{125}I labeled DNA

350 nucleotide pea DNA was incubated to Cot 400 and the single stranded fraction (15%) isolated on hydroxyapatite. This material was concentrated and labeled with ^{125}I using a modified Commerford (1971) procedure (Holmes and Bonner, 1974).

Renaturation of DNA

DNA fragments were denatured at 100 °C and renatured in 0.12 M sodium phosphate buffer (PB) at 60 °C and 70 °C and in 0.48 M PB at 71 °C and 81 °C. Samples were then diluted to 0.03 M PB and passed over hydroxyapatite at 60 °C to check for degradation. Samples with more than 10% of the DNA not binding to hydroxyapatite in 0.03 M PB were discarded. The single strand fraction was then eluted with 0.12 M PB and the double strand fraction denatured at 100 °C and eluted with 0.12 M PB.

Optical renaturation was carried out in a Gilford 2400 spectrophotometer equipped with a water jacketed sample compartment. Samples were denatured, loaded into the sample compartment and the decrease in hyperchromitcity monitored. The data in Figure 2A were measured in 0.12 M PB at 60 ^{o}C in 10 mm and 1 mm path length cuvettes and in 0.41 M PB at 71 ^{o}C in 10 mm cuvettes. The data in Figure 2B was measured at 66 ^{o}C in 10 mm and 1 mm cuvettes.

Hybridization of ^{125}I labeled DNA was carried out in 0.12 M PB at 60 ^{o}C and in 0.48 M PB at 65 ^{o}C. After renaturation, samples were diluted to 0.12 M PB and passed over hydroxyapatite in 0.12 M PB. The double strand fraction was eluted with 0.48 M PB. The renaturation of the driver DNA was followed optically; the fraction of the ^{125}I tracer in hybrid was measured by TCA precipitation of the hydroxyapatite fractions.

Fits of second order components to the data were calculated using a non-linear least squares computer program (Pearson et. al., 1977).

Results

To determine the fraction of single copy and the single copy rate constant we have made three types of measurements. We have renatured pea DNA at 60 ^{o}C and 70 ^{o}C assaying the fraction duplex by hydroxapatite (HAP) binding. We have optically renatured pea DNA at 60 ^{o}C and 66 ^{o}C to determine the true fraction duplex. And we have driven an isolated labeled slowly renaturing fraction to look for a small fraction highly complex component. Renaturations were carried out at two criterion

temperatures to look for apparent repetitive fractions caused by poorly matched sequences. Both hydroxyapatite and optical Cot curves were used to discover if an apparent increased repetitive fraction might be due to repeated sequences interspersed with short single copy spacers.

The hydroxyapatite Cot curves

Figures 1A and 1B show the renaturation of 350 nucleotide long pea DNA fragments at 60 °C and 70 °C. The 60 °C (T_m-25°C) incubations were carried out at 60 °C in .12 M sodium phosphate buffer (PB, .18 M Na^+) and at 71 °C for .41 M PB. The 70 °C (T_m-15°C) curve was carried out at 70° for .12 M PB and 81 °C for .41 M PB. The higher temperatures were used to correct for the higher melting temperature of DNA in .41 M salt. These high temperatures may account for some of the scatter in the data and the final fraction unreacted at high Cot's.

Tables IA and IB present the results of a variety of least squares fits to the data. The tables start with the results of an unconstrained two component fit. This is the "best fit" possible for the data. The second fits show the results when the single copy rate from the renaturation at one temperature is forced on the Cot curve at the other temperature. This calculation gives another measure of the real difference between the slow component rates at the different temperatures. These fits suggest the difference between the 60 °C and 70 °C Cot curves is not significant. The apparent shift of some repeated sequences to the slow component increases the rate of the slow component and may be due to the difficulty of resolving components separated by two decades.

Figure 1: Renaturation of pea DNA measured by hydroxyapatite binding

Denatured 350 nucleotide pea DNA was renatured in 0.12 M PB and 0.41 M PB at 60 °C and 71°C (T_m-25 °C) or 70 °C and 81 °C (T_m-15 °C) respectively. The fraction single stranded was determined by hydroxyapatite chromatography. Equivalent Cot is the Cot (moles/liter-sec) times salt concentration factors of 1.0 for 0.12 M PB incubations and 5.0 for 0.41 M PB incubations (Britten et. al., 1974).

The solid lines plotted show the best least squares fit to the data. The dashed line shows the contribution of a component with a rate of 0.0004 liter-sec/mole.

A. Renaturation at T_m-25 °C.

B. Renaturation at T_m-15 °C.

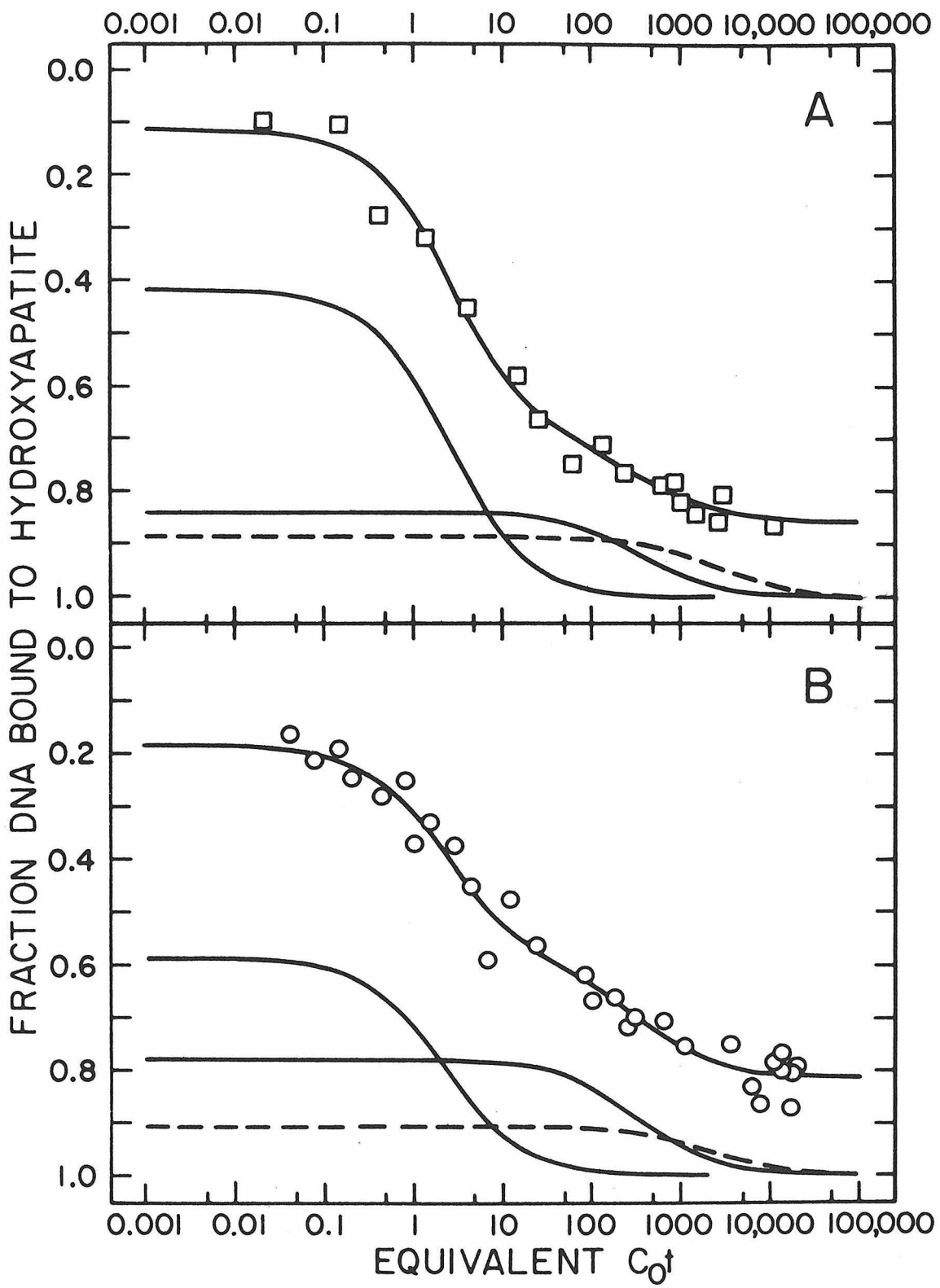

0.001 0.01 0.1 1 10 100 1000 10,000 100,000
0.0
0.2
0.4
0.6
0.8
1.0
A
B
FRACTION DNA BOUND TO HYDROXYAPATITE
EQUIVALENT C_0t

Table I

Renaturation of 350 nucleotide pea DNA fragments at 60 °C and 70 °C

Renaturation measured by hydroxyapatite binding

A. Unconstrained fits: (no parameters fixed)

1. 60 °C (T_m-25°C) reaction

Goodness of fit: 3.7%

Component	Fraction	Rate	Cot 1/2	Repetition Frequency
1	0.114	> 50	0.02	>20,000
2	0.583 (±0.046)	0.399 (±0.128)	2.51	138
3	0.160 (±0.043)	0.00288 (±0.00315)	347	1
	Final fraction unreacted:		0.143 ±0.029	

2. 70 °C (T_m-15°C) reaction

Goodness of fit: 4.0%

1	0.1840	> 50	0.02	>20,000
2	0.411 (±0.043)	0.452 (±0.174)	2.21	148
3	0.220 (±0.044)	0.00305 (±0.00202)	3279	1
	Final fraction unreacted:		0.185 ±0.015	

B. Constrained fit: single copy rate from reaction at one temperature used to fit reaction at the other temperature

1. 60 °C (T_m-25°C) reaction

Goodness of fit: 3.5%

1	0.113	> 50	0.02	>20,000
2	0.581 (±0.034)	0.403 (±0.108)	2.48	132
3	0.161 (±0.039)	0.00305	328	1

Final fraction unreacted: 0.145 ±0.018

1. 70 °C (T_m-15°C) reaction

Goodness of fit: 3.9%

1	0.184	> 50	0.02	>20,000
2	0.414 (±0.029)	0.443 (±0.130)	2.25	154
3	0.217 (±0.031)	0.00288	347	1

Final fraction unreacted: 0.185 ±0.013

C. Constrained fit: three fitted components single copy component with rate 0.0004

1. 60 °C (T_m-25°C) reaction

Goodness of fit: 3.1%

1	0.0650	> 50	0.02	>125,000
2	0.282 (±0.108)	2.74 (±2.79)	0.365	6850
3	0.428 (±0.111)	0.100 (±0.057)	10	250
4	0.116 (±0.050)	0.00040	2500	1

Final fraction unreacted: 0.109 ±0.033

2. 70 °C (T_m-15°C) reaction

Goodness of fit: 4.1%

1	0 .179	> 50	0.02	>125,000
2	0.382 (±0.070)	0.547 (±0.296)	1.83	1367
3	0.172 (±0.073)	0.0103 (±0.0166)	97	26
4	0.096 (±0.084)	0.00040	25000	1

Final fraction unreacted: 0.171 ±0.022

A third set of fits is included to test for a small fraction of sequences hybridizing at the rate of 0.0004 liter-sec/mole with a Cot 1/2 of 2500. These fits indicate our measurements cannot exclude the possibility of a very slowly renaturing fraction. The dashed lines in Figures 1A and 1B show how this hypothetical component would renature.

From both of these measurements it appears that about 60% of pea DNA is repetitive with a Cot 1/2 of 1 and another 25% is less repetitive, with a Cot 1/2 of 350 corresponding to a genome size of .35 pg. There is not a substantial difference between the renaturation at 60 ^{o}C and at 70 ^{o}C.

The optical Cot curves

Hydroxyapatite fractionation measures the formation of duplex containing complexes. Strands containing duplexes longer than 20-40 nucleotides are retained on HAP while a large fraction of the retained strands may be single stranded. Repeated sequences 300 nucleotides long separated by single copy sequences of equal length would cause the amount of repetitive HAP binding to be twice the actual amount of repetitive sequences.

Optical renaturation - measuring the fraction duplex by renaturing the DNA in a spectrophotometer - measures the true duplex fraction in the DNA. Single strand ends of duplexed molecules contribute no more to the hyperchomicity measurement than unduplexed single strand molecules.

Figure 2: Optical renaturation of pea DNA

350 nucleotide pea DNA fragments were denatured and renatured in a water jacketed spectrophotometer chamber. Samples were incubated in 0.12 M PB at 0.41 M PB at 60 ^{o}C and 71 ^{o}C (T_m-25 ^{o}C) in 0.12 M PB at 66 ^{o}C. Fraction DNA reacted is the fraction of hyperchomicity lost after renaturation. Equivalent Cot was calculated as in Figure 1.

The solid lines plotted show the results of the best least squares fits to the data. The dashed lines show the contribution of a component renaturing with a rate of 0.0002 liter-sec/mole.

A. Renaturation at T_m-25 ^{o}C.

B. Renaturation at T_m-19 ^{o}C.

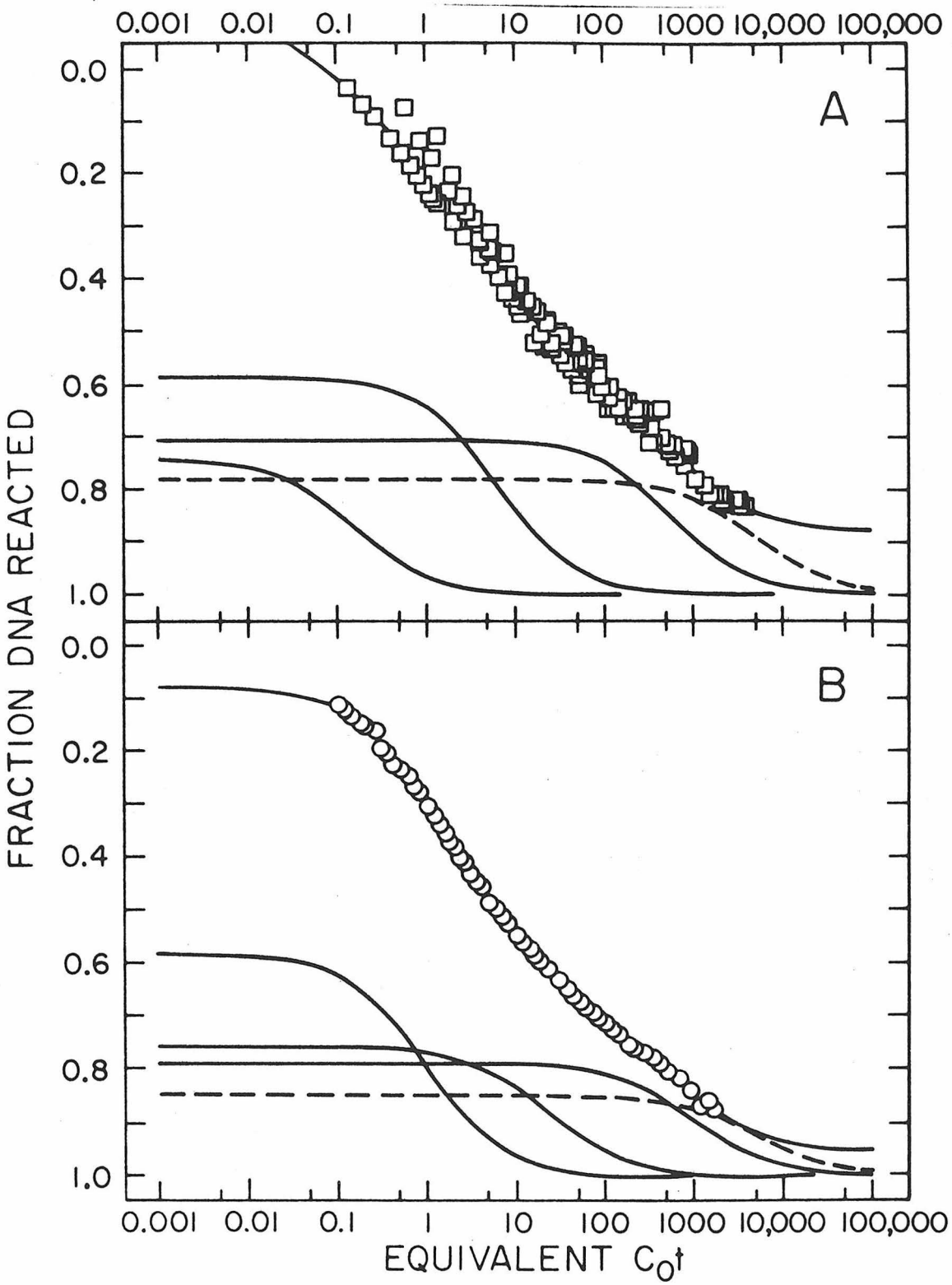

0.001
0.01
0.1
1
10
100
1000
10,000
100,000
0.0
0.2
0.4
0.6
0.8
1.0
A
B
FRACTION DNA REACTED
EQUIVALENT C_0t

Table II

Renaturation of pea DNA fragments at 60 °C and 66 °C

Renaturation measured optically

A. Unconstrained fit: (no parameters fixed)

1. 60 °C (T_m-25 °C) reaction

Goodness of fit: 2.6%

Component	Fraction	Rate	Cot 1/2	repetition frequency
1	0.261 (±0.181)	6.70 (±12.9)	0.15	3800.
2	0.414 (±0.051)	0.167 (±0.046)	6.0	95.
3	0.295 (±0.016)	0.00176 (±0.00040)	568.2	1.

Final fraction unreacted: 0.122 ±0.017

2. 66 °C (T_m-19 °C) reaction

Goodness of fit: 0.5%

1	0.071	>100	0.01	>100,000
2	0.423 (±0.0112)	1.08 (±0.078)	0.926	1030.
3	0.246 (±0.010)	0.0467 (±0.0075)	21.41	45.
4	0.214 (±0.0.15)	0.00104 (±0.00031)	961.5	1.

Final fraction unreacted: 0.046 ±0.022

B. Constrained fit: (single copy rate fixed at 0.0002 liter-sec/mole)

1. 60 °C (T_m-25 °C) reaction

Goodness of fit: 2.7%

1	0.067	>10	<0.1	>5000.
2	0.465 (±0.017)	0.323 (±0.046)	3.10	1600.
3	0.229 (±0.016)	0.00551 (±0.00163)	181.5	27.6
4	0.219 (±0.016)	<u>0.0002</u>	5000.	1.

Final fraction unreacted: <u>0.020</u>[a]

2. 66 °C (T_m-19 °C) reaction

Goodness of fit: 0.8%

1	0.087	<0.01	>100.	>50,000.
2	0.482 (±0.007)	0.752 (±0.042)	1.33	3760.
3	0.259 (±0.007)	0.0134 (±0.0014)	75.	66.7
4	0.152 (±0.005)	<u>0.0002</u>	5000.	1.

Final fraction unreacted: <u>0.020</u>[a]

Footnotes to Table II

[a] These values were fixed to allow the program to do the fit.

Figures 2A and 2B display the renaturation of pea DNA at 60 °C and 66 °C in the spectrophotometer. These curves show the same large repetitive fraction seen in the HAP Cot curves. In these curves it is about 45%. Again a number of multicomponent curves were fit to the data. Tables IIA and IIB show the results of those fits. The first fit for each set of data is the best two component fit. We also include a three component fit. The third fit forces the slow component rate from one renaturation temperature on the data of the other. The final curve forces a slow component appropriate for a 2.5 pg genome size.

A comparison of these data with the HAP cot curves indicates that perhaps 30% of the 60% repetitive fraction from the HAP curve may be due to single strand tails. The rate constants for the two components are not significantly different.

Hybridization of labeled slowly renaturing DNA

Each of the renaturation curves we have run suggest there are only two major components in the pea genome. The slow component reanneals quite rapidly, about 20 times faster than that expected from the 5 pg. genome size predicted from a nuclear DNA content of 9.7 pg (Birnsteil et. al., 1964). Because of the poor termination data and small fraction of the DNA in the complex component it is difficult to rule out a minor component of with a Cot 1/2 of 2500.

We have increased the sensitivity of our measurement more than four fold by isolating a slowly renaturing fraction of the genome, labeling it <u>in vitro</u> with ^{125}I and driving that tracer with whole pea DNA. Pea DNA was renatured to Cot 400 and the single strand fraction separated on

Figure 3: Hybridization of ^{125}I labeled single copy pea DNA

A fraction of pea DNA (15%) enriched in single copy sequences was isolated and labeled as described in the text. This labeled fraction was driven by a 5000 fold excess of unfractionated pea DNA to the equivalent Cots shown. The renaturation of the driver DNA (not shown) and ^{125}I single copy tracer was measured after fractionation on hydroxyapatite. The solid line shows the best least squares fit with the components described in Table III.

(■) fraction of ^{125}I DNA in duplex.

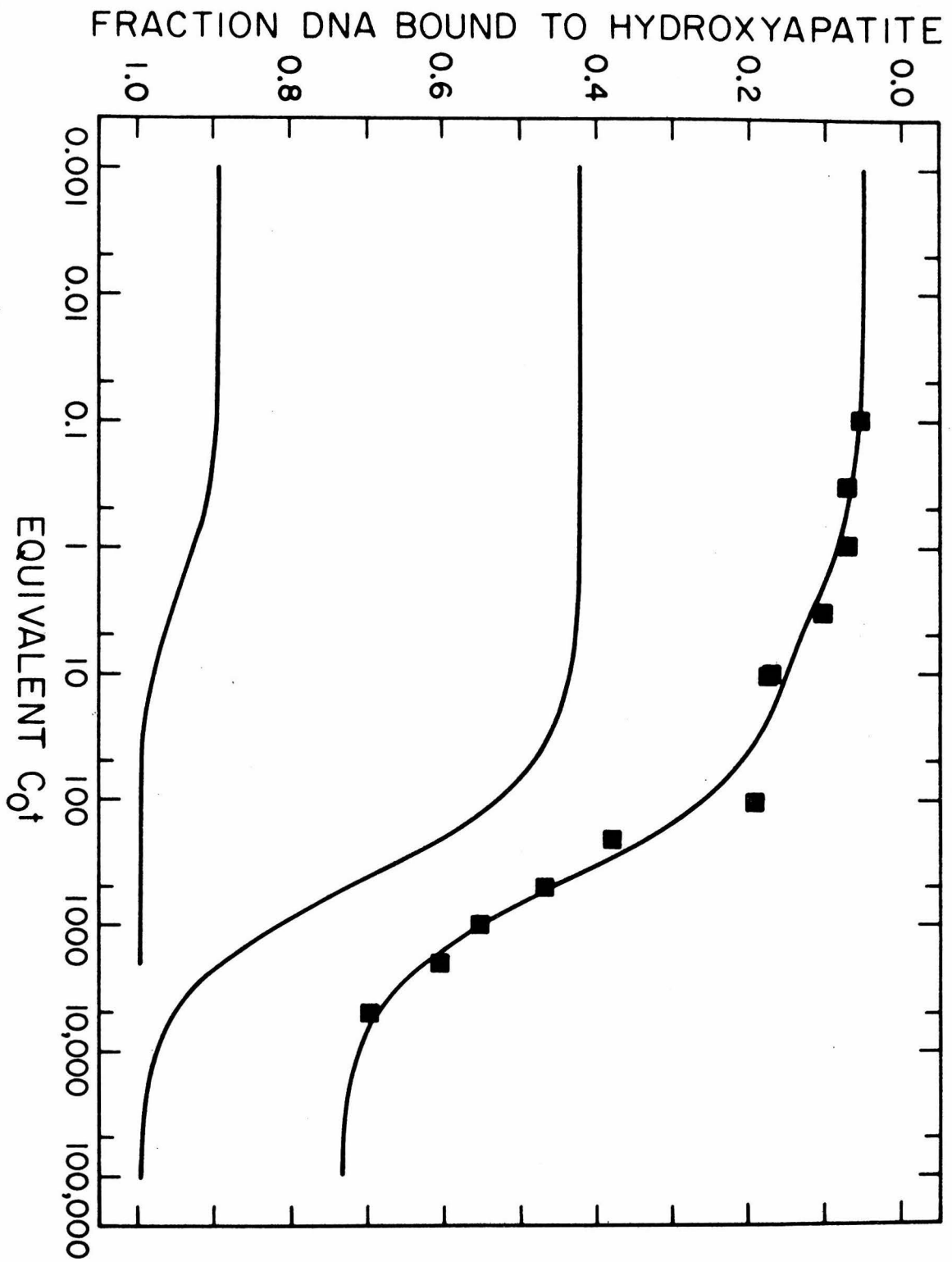

FRACTION DNA BOUND TO HYDROXYAPATITE
1.0
0.8
0.6
0.4
0.2
0.0
EQUIVALENT C_0t
0.001
0.01
0.1
1
10
100
1000
10,000
100,000

Table III

Hybridization of ^{125}I labeled

Slowly renaturing pea DNA

Goodness of fit: 3.6%

Component	Fraction	Rate	Cot 1/2	repetition frequency
1	0.109 (±0.051)	0.463 (±0.972)	2.16	215.
2	0.580 (±0.053)	0.00215 (±0.00082)	465.	1.

Final fraction unreacted: 0.263 ±0.044

HAP. This material was labeled with ^{125}I using the Commerford (1971, Holmes and Bonner, 1974) procedure and driven by whole unlabeled DNA. Fifteen percent of the DNA was isolated and labeled for this experiment so a 10% single copy fraction in whole DNA should appear as more than 50% of the DNA after the purification. The hybridization curve in Figure 3 does not indicate any component with a Cot 1/2 greater than 500.

Discussion

These measurements are all in good agreement as to the fraction of the genome in the repetitive component. The HAP Cot curves suggest that about 60% of the DNA is repetitive while the optical Cot curves indicate that slightly less than 45% is repetitive.

Both curves show the same two kinetic components. The fast component is centered at Cot 1, the slow component around Cot 400. The hybridization of the isolated slow component DNA driven by unfractionated pea DNA shows the same slow component.

The conclusion that about 45% or more of the DNA sequences in the pea genome are repeated is a strong one. Changes in the incubation temperature which should melt poorly matched duplexes do not alter the renaturation curve. The large fraction repetitive is really due to duplexed DNA, not to single strand tails or closely spaced duplexes. We are therefore confident that the fraction of repeated sequences we have determined is correct.

There is also good agreement on the Cot 1/2 of the slowly renaturing complex component - about 450. If this slow component represents the single copy DNA of the pea the genome size is about .46 pg. Unfortunately it is difficult to be certain that the final fraction unreacted does not contain a more slowly renaturing component. Ten percent of the DNA could easily be renaturing but hidden in the final unreacted fraction of either HAP Cot curve or the optical Cot curves. We believe it unlikely that a more slowly reannealing component is to be found in the pea genome. The hybridization experiment displayed in Figure 3 sets an upper limit on the fraction of very slow component below 5% of the pea genome.

We believe that the true size of the pea genome is 5×10^8 base pairs. This is in contrast to the 5×10^9 base pairs/genome based on the chemical determination of DNA in nuclei isolated from pea seedling stem tips (Birnsteil et. al., 1964) or root tips (McLeisch and Sunderland, 1961; Van't Hoff and Sparrow, 1963). We now know (Libbenga and Torrey, 1973) that the majority of stem cells of the pea are 4X to 16X polyploid. Walbot and Dure (1975) made a similar discovery on measuring the kinetics of reassociation of cotton DNA. We believe that the complexity measured from the rate of renaturation of single copy DNA (Britten and Kohne, 1968; Laird, 1971) provides a more accurate value for the genome size.

References cited

Birnsteil, M. L., Chipchase, M. I. H. and Flamm, W. G. (1964) "On the chemistry and organization of nucleolar proteins" Biochem. Biophys. Acta. 87:111-122

Britten, R. J., Graham, D. E. and Neufeld, B. R. (1974) "Analysis of repeating DNA sequences by reassociation" Methods in Enzymology (L. Grossman and K. Moldave, eds.) Vol 29 part E 363-418

Britten, R. J. and Kohne, D. E. (1968) "Repeated sequences in DNA" Science 161:529-540

Commerford, S. L. (1971) "Iodination of nucleic acids in vitro" Biochemistry 10:1993-2000

Davidson, E. H., Galau, G. A., Angerer, R. C. and Britten, R. J. (1975) "Comparative aspects of DNA sequence organization in metazoa" Chromosoma (Berl.) 51:253-259

Holmes, D. S. and Bonner, J. (1974) "Sequence composition of rat nuclear deoxyribonucleic acid and high molecular weight ribonucleic acid" Biochemistry 13:841-848

Laird, C. D. (1971) "Chromatid structure: relationship between DNA content and nucleotide sequence diversity" Chromosoma (Berl.) 32:378-406

Libbenga, K. R. and Torrey, J. G. (1973) "Hormone induced endoreduplication prior to mitosis in cultured pea root cortex cells" American Journal of Botany 60:293-299

McLeisch, J. and Sunderland, N. (1961) "Measurements on desoxyribosenucleicacid (DNA) in higher plants by Feulgen photometry and chemical methods" Expt. Cell Res. 24:527-540

Pearson, W. R., Davidson, E. H. and Britten, R. J. (1977) "A program for least squares analysis of reassociation and hybridization data" submitted to Nuc. Acids Res.

Smith, D. B., Rimpau, J. and Flavell, R. B. (1976) "Interspersion of different repeated sequences in the wheat genome revealed by interspecies DNA/DNA hybridization" Nuc. Acids Res. 10:2811-2825

Van't Hoff, J. and Sparrow, A. H. (1963) "A relationship between DNA content, nuclear volume and minimum mitotic cycle time" Proc. Nat. Acad. Sci. (US) 49:897-902

Walbot, V. and Dure, L. S. (1976) "Developmental biochemistry of cotton seed embryogenesis and germination. VII. Characterization of the cotton genome" J. Mol. Biol. 101:503-536

APPENDIX A

A PROGRAM FOR LEAST SQUARES ANALYSIS OF REASSOCIATION AND HYBRIDIZATION DATA

Abstract

A computer program is described for the rapid calculation of least squares solutions for data fitted to different functions normally used in reassociation and hybridization kinetic measurements. The equations for the fraction not reacted as a function of Cot follow: First order, $\exp(-kCot)$; second order, $(1+kCot)^{-1}$; variable order, $(1+kCot)^{-n}$; approximate fraction of DNA sequence remaining single stranded, $(1+kCot)^{-.44}$; and a function describing the pairing of tracer when the rate constant for the tracer (k) is distinct from the driver rate constant (k_d): $\exp\{k[1-(1+k_dCot)^{1-n}]/[k_d(1-n)]\}$. Several components may be used for most of these functional forms. The standard deviations of the individual parameters at the solutions are calculated.

Introduction

The quantitative examination of reassociation and hybridization kinetic measurements has become increasingly important as the sophistication of the measurements has grown. In this paper we describe a general computer program which can conveniently apply the variety of functions now used to interpret kinetic measurements. We have chosen to use a non-linear least squares method so that the solutions give equal weight to all of the individual measurements and no preliminary assumptions need be made about the initial or terminal values of the reaction.

Least squares computer programs have been applied to the problem of resolving repetitive and single copy kinetic components in DNA renaturation experiments carried out on many organisms (e.g. Davidson and Britten, 1973; Davidson et. al., 1975a). They are also used to determine rate constants in RNA excess hybridization reactions (e.g. Galau et. al., 1975). The use of cDNA probes to determine the complexity of RNA populations by kinetic rather than saturation measurements (e.g. Bishop et. al., 1974; Ryffel and McCarthy, 1975) also relies heavily on the resolution of abundance classes by accurate determination of their rate constants.

The five functions listed in Table 1 are used for the examination of the following kinds of measurements. The second order equation (FINGER) describes accurately the form of DNA reassociation kinetics assayed by hydroxyapatite chromatography (Britten and Kohne, 1966; Wetmur and Davidson, 1968) though for fairly complex reasons (Britten and Davidson, 1976; Smith et. al., 1975). The pseudo-first- order

TABLE 1 Functional forms used by the program

Number	Name	Form	Use
1	FINGER	$f_i(1 + k_i Cot)^{-1}$	second order reaction: DNA renaturation measured by hydroxyapatite chromatography
2	WHATOR	$f(1 + kCot)^{-n}$	variable form reaction: to determine values of n when the apparent order of the reaction is unknown
3	NUFORM	$f_i \exp\{k_i \lfloor 1-(k_d Cot)^{1-n} \rfloor / \lfloor k_d(1-n) \rfloor\}$	describes rate of tracer reaction when tracer rate constant k differs from driver rate constant k_d
4	EXCESS	$f_i \exp(-k_i Cot)$	first order function: for RNA excess experiments
5	WHTCMP	$f_i(1 + k_i Cot)^{-0.44}$	modified second order function: S-1 nuclease assay of hybridization

equation (EXCESS) applies when the nucleic acid driving the reaction remains unpaired as in RNA driven reactions. The third function (WHATOR) has a form which can be varied by changing the value of the exponent, and is useful when the apparent order of the reaction is not known or there is a need to test the heterogeneity of the reacting components. When the exponent n=1, it reduces to second order form. When high quality single component hydroxyapatite measurements are analyzed, the least squares solution yields n=1.0 (Smith et. al., 1975). As n becomes large this equation approaches first order and values of n as high as 10 yield a form which is indistinguishable from first order. The next functional form (WHTCMP) is applicable to DNA reassociation when the reaction is assayed by S-1 nuclease (Smith et. al., 1975) and expresses the fraction of the length of the DNA sequence present which remains single stranded. The last equation (NUFORM) applies when the rate of reassociation of tracer with the driver and the rate of driver renaturation itself differ. The amount of driver available is assumed to follow the equation $(1+kCot)^{-n}$. A value of n=.44 is usually used and gives a good approximation to the actual capacity of the remaining single stranded regions to reassociate with tracer molecules (Davidson et. al., 1975b).

The program: Use

The program presented in Appendix III is in use on a PDP-10 timesharing computer. It provides for interactive input but has been designed for ease of conversion to a Batch processing system. A typical run is presented in Appendix I, with user responses underlined. In a Batch environment the underlined inputs would be submitted on cards.

Upon execution in a timesharing environment, the program types a list of options available:

1START,2ADD,3GUESS,5PLOT,6CPLOT,8LPLOT,9STOP

Options 1 and 2 are for input, option 3 is used to find the solution, and options 5, 6 and 8 are used for output. Option 9 terminates the program.

Input options: (1START,2ADD)

1START is used at the beginning of the program to read in a new set of data. The program asks for a data file name which should contain input data consisting of:

1. Title of the data file
2. C/Co (the fraction of DNA or RNA remaining single stranded), and the value of the Cot for each data point
3. 3.0 in the C/Co position to terminate input

The form of these data is described more fully in the program listing.

Option 2 permits addition of a new file to a pre-existing set of data. This program handles up to 200 data points.

The fitting option: (3GUESS)

The user is requested to provide the following three lines of information:

1. FUNCTION(-1HELP),ITERATIONS,RMS QUIT,DELMX QUIT,(NIT)

2. parameter estimates

3. index of parameters to be fit

First line:

FUNCTION may be given the following values (see Table I):

-1. (HELP) prints out the names and numbers of the functions available.

OSAME,1FINGER,2WHATOR,3NUFORM,4EXCESS,5WHTCMP

0. (SAME) The previously used function, parameter values and fixed parameters are saved to make a new approximation

1. set of second order components (FINGER)

$$\langle C/Co \rangle = F + \sum_{i=1}^{r} f_i (1+k_i Cot)^{-1} \quad r<=5$$

2. variable exponent form to examine the order or heterogeneity of a reaction (WHATOR)

$$\langle C/Co \rangle = F + f(1+kCot)^{-n}$$

3. tracer reaction with driver of different rate constant (NUFORM)

$$\langle C/Co \rangle = F + \sum_{i=1}^{r} f_i \exp\{k_i [1-(1+k_d Cot)^{1-n}]/[k_d(1-n)]\} \quad r<=2$$

4. set of first order components (EXCESS)

$$\langle C/Co \rangle = F + \sum_{i=1}^{r} f_i \exp(-k_i Rot) \quad r<=5$$

5. set of components for reassociation assayed with a single strand specific enzyme (WHTCMP) $\langle C/Co \rangle = F + \sum_{i=1}^{r} f_i (1+k_i Cot)^{-.44}$ $r<=5$

In the equations <C/Co> is the calculated fraction of the DNA which remains single stranded according to the assay involved. F is the fraction which is not reacted at the termination of the reaction of the components described. f_i is the fraction in an individual component and k_i is the rate constant. f and k are these parameters for a single component case. k_d is the rate constant for self-reassociation of the driver DNA. n is the variable value of the exponent where required.

ITERATIONS is the number of iterations to be used to fit the function to the data. For reasonable guesses 5 iterations are sufficient. If the fit has not converged after 20 iterations, it will probably not converge. These situations are discussed in the section on PROBLEMS. To calculate the error estimates for an arbitrary set of parameters without iterating, 0 ITERATIONS should be typed.

RMS QUIT and DELMX QUIT provide additional criteria for terminating the data fitting process. The process will terminate either when the difference between the RMS and the current and previous iterations is less than RMS QUIT or when the DELMX value (the largest fractional change in any parameter) is less than DELMX QUIT. If RMS QUIT or DELMX QUIT is zero, the process will terminate when either the RMS or the parameters are not changing. (NIT) is a value for a corrective stragtegy used in the program. It is set to 10 in the program by default, and in general should not be used.

Second line:

The initial parameter guesses are to be specified by the user in the following order: F, f1,k1, f2,k2, f3,k3, f4,k4, f5,k5 for functions

1(FINGER), 4(EXCESS) and 5(WHTCMP). For function 2(WHATOR) the order is F, f, k, n. For function 3(NUFORM) the order is F, f1,k1, f2,k2, k_d, n.

Third line:

The index or order number of the parameters which are to be fixed at their initial values is specified by the user. Any parameter or number of parameters may be fixed.

After these specifications are supplied the iterations are performed. The RMS, DELMX (the maximum fractional change in any parameter) and the parameter values are printed after each iteration until the ITERATIONS are finished or one of the termination conditions specified by RMS QUIT or DELMX QUIT is satisfied.

When the iterations stop, the parameter correlation matrix and two criteria for the error of the fit, the RMS and the GOODNESS OF FIT are printed, as well as the NUMBER OF data POINTS and the number of DEGREES OF FREEDOM. The number of DEGREES OF FREEDOM is the NUMBER OF POINTS less the number of variable (not fixed) parameters. The RMS and GOODNESS OF FIT values are calculated from the sum of the squared errors.

$$SUMSQ = \sum (C/Co - \langle C/Co \rangle)^2$$

for each of the data points. C/Co is the fraction unreacted at the Cot of the data point and <C/Co> is the expected value calculated from the function used to fit the data. The RMS value printed then is:

$$RMS = SUMSQ/p$$

where p is the number of data points. The GOODNESS OF FIT value uses the DEGREES OF FREEDOM instead of the number of points:

GOODNESS OF FIT = $\sqrt{SUMSQ/(p-v)}$

where v is the number of variable parameters in the fit.

Output options: (5PLOT,6CPLOT,8LPLOT)

Options 5 and 8 print an alphanumeric plot of the solution and the data on the teletype or line printer respectively. Option 6 causes a file to be written which can be used for plotting by another program (such as one which uses an X-Y plotter).

For example, the following sequence of responses could be used to fit a second order function to the data in the file named FILE and have the result plotted on the line printer (see Appendix I).

```
1.  1       ;start

2.  FILE    ;name of a PDP-10 data file
            ;(this statement would probably be deleted
            ; in a batch system)

3.  1       ;type out the data

4.  3       ;(GUESS) do the fit

5.  1 10 0.0 .005
            ;fit a second order function (FINGER)
            ;use 10 ITERATIONS
            ;terminate if RMS is not changing
            ;terminate if parameters are changing
            ; by 0.5% or less

6.  0.05 0.25 10.0 0.65 0.001
            ;parameter estimates

7.  5       ;the rate constant of the second
            ; component is fixed

8.  8       ;plot the fit on line printer

9.  9       ;stop the program
```

A typical run using the program is presented in Appendix I. It shows a number of the problems encountered when fitting hybridization data which are discussed in the next section. Appendix II presents a discussion of the structure and mathematical techniques used in the program. It is not necessary to read Appendix II to use the program; the material is directed to those who wish to modify the program or apply it to a different problem.

Appendix III presents a highly annotated listing of the program. We have made efforts to provide a program that will run on different computers but some statements (particularly input statements) may have to be changed when the program is used on other systems. The program listing describes in some detail how these changes might be carried out.

Analysis of hybridization data

The program strategy

This non-linear least squares program is designed to converge on a solution yielding parameters which minimize the least squares deviation of the function from a set of data. There are two main concerns: whether the final solution is biased by the input parameter estimates, and whether further iterations will improve the solution. A solution is independent of input parameter estimates and insensitive to further iterations if it has converged. Convergence is indicated by the amount of change in the RMS from one iteration to the next and by the change in the DELMX parameter. Because of the Taylor's series approximation technique used by this program, the parameter values converge very rapidly in the neighborhood of a solution. Although this neighborhood

may be difficult to find for the first few iterations (the algorithm's strategy uses a "gradient" technique to find the neighborhood), once found, successive iterations will improve the parameter values by one to two significant figures with each iteration. This improvement is reflected in a 10- to 100- fold decrease in DELMX with successive iterations. This rapid decrease in DELMX may occur while the RMS changes very little. Attempts to improve the solution beyond convergence may cause the message CORRECTIVE ITERATIONS EXCEEDED to be printed.

Possible problems

This program rapidly converges to a unique solution from a wide range of parameter estimates when the data provide adequate constraints. New parameter estimates are always better (in the least squares sense) than the previous values but the parameters may become negative during the process. Inadequate termination data may allow the FINAL parameter to become negative while the FRACTION parameter for the slowest component increases without bound. Fixing the FINAL fraction unreassociated (presumably near zero) or close scrutiny of the DELMX values to find regions of local convergence may solve the problem.

Often when the series of iterations do not immediately converge the program hesitates in a region of local convergence. When pressed to improve the RMS the program may go off in a direction (such as negative parameters) where termination is impossible. Usually DELMX starts at 1.0 to 5.0, then decreases to 0.02 to 0.1 as the fit converges (the largest parameter change is less than 2-10%). DELMX may then jump by a factor of 50 to 200 and start on another (possibly unterminated) path.

The small DELMX value indicates a local low RMS region which might provide good values for the parameters but they must be carefully examined. To determine the quality of the parameter values, plot the solution.

Occasionally the program returns negative rate constants for components of the reaction. This is usually due to an attempt by the program to remove that component from the solution. Plotting the solution usually shows why this was done. Single erroneous points that do not show in the plot should be searched for and perhaps a new solution should be attempted with fewer components.

Parameter fixing

Data from other measurements may supply fixed values for parameters in a solution. For example, chemical measurements of genome size may be used to establish the rate constant of the single copy component. Slave "mini-cot" experiments (Britten et. al., 1974) often provide the most reliable determinations of repetitive and single copy rate constants which may be used as fixed values to calculate the fraction of the genome associated with the different rate components in a total reassociation curve.

Fixing the single copy rate constant at a value determined from the genome size or minicot analysis may be particularly important for DNA which contains a small fraction repeated 2-20 fold. This low repetition class is virtually indistinguishable from single copy DNA but can increase the apparent single copy rate constant by a factor of two.

Parameter fixing also provides information about the variety of similar least squares solutions which describe the data equally well. A graph of RMS vs. a set of fixed values of a parameter can be informative, particularly if it turns out to be a very shallow curve near the minimum.

Parameter statistics

Parameter standard deviations and correlation coefficients calculated by the program provide information about the range of parameter values which may adequately describe the data. These numbers, along with the RMS and DELMX, indicate the significance of the parameter values obtained at the least squares solution.

The parameter standard deviation and correlation coefficient calculation is similar to standard deviation and correlation coefficient calculations in linear regression analysis. To obtain the standard deviations and correlation coefficients the data covariance matrix is inverted. The calculation assumes that the Taylor's series linear approximation is accurate in the neighborhood of the solution (indicated by a low DELMX at convergence). To examine the usefulness of the parameter standard deviations and correlation coefficients, artificial test data were generated. Some examples of solutions using the second order (FINGER) function are shown in Table 2.

The data set for each of these analyses was calculated using a Gaussian random number generator for the final unreacted quantity as follows:

$$C/Co = GAUS(F,FSD) + \sum_{i=1}^{a} f_i(1+k_i Cot)^{-1}$$

TABLE 2 EFFECT OF GAUSSIAN NOISE ON PARAMETER VALUES AND ERROR ESTIMATES

Successive trials with random noise

	Trial	Number of points	Parameter values at solution					RMS
			Final	f_1	k_1	f_2	k_2	
Value	Parameters used		0.200	0.300	0.1000	0.300	10.00	
Error	to generate data		0.040[a]					
Solution	1	40	0.196	0.235	0.0536	0.328	3.69	0.0350
Error			0.017	0.037	0.0289	0.038	1.38	
Solution	2	40	0.198	0.309	0.106	0.297	13.5	0.0466
Error			0.019	0.038	0.048	0.038	6.6	
Solution	3	40	0.212	0.301	0.149	0.319	18.9	0.0392
Error			0.015	0.032	0.058	0.032	7.4	
Solution	4	40	0.208	0.274	0.0819	0.300	5.82	0.0382
Error			0.017	0.040	0.0397	0.041	2.47	
Solution	5	40	0.180	0.340	0.110	0.302	10.2	0.0379
Error			0.015	0.035	0.039	0.035	4.2	
Solution	sum of	200	0.200	0.298	0.103	0.300	9.28	0.0403
Error	trials above		0.007	0.017	0.020	0.017	1.81	

[a] Noise factor equal to FSD described in the text. For this example, F is 0.2

(GAUS is a computer subroutine which generates a series of values following a Gaussian distribution with average F and standard deviation FSD.)

This method of generating variation in the "data" is a good analogue of actual fluctuations from measurement to measurement, such as the binding of DNA to hydroxyapatite. The fluctuations observed in Table 2 for the various solutions and the standard deviations shown are probably representative of what would happen in repetitions of actual measurements which showed a comparable RMS. It is clear that the standard deviations of the rate constants are much larger than those of the component quantities. When analyses of this sort are carried out with components only a factor of ten apart in rate constant (instead of the factor of 100 for the example illustrated in Table 2) the rate constant fluctations are very large. It can be seen from Table 2 that there is a reasonable quantitative relationship between the fluctuations from set to set and the calculated standard deviations.

This table exhibits one of the characteristic problems of fitting reassociation kinetic data and indicates the insight into the accuracy of individual parameter estimates provided by the standard deviation calculations. It is clear that where two kinetic components are present, even a factor of one hundred apart, very accurate data are needed to obtain estimates of rate constants with moderate accuracy.

Discussion

We have described a powerful and flexible program for the least squares analysis of the kinetics of reassociation. This program has been used for the analysis of an extensive series of measurements (Galau et. al., 1974; Davidson et. al., 1975b; Smith et. al., 1975; Britten and Davidson, 1976; Galau et. al., 1976). Least squares analysis is necessary for reproducible and clear interpretation of such measurements. We refer to these papers for examples of use and interpretation, while in this discussion we focus on problems of over-interpretation or misinterpretation of the solutions. As powerful analytical tools of this sort are developed it becomes very easy to simply accept the "output" and lose touch with its meaning.

It is important to remember that any successful least squares strategy provides parameter values which "fit the data better" in the "least squares sense". The solution does not guarantee either physical or biological reality. Components may be used to fit small peculiarities in the data and have little physical meaning. This is usually evident from a plot of the fit. Parameter values may also be the fluke of a particular set of data and have little absolute importance. The parameter standard deviations calculated by the program provide confidence limits for the parameter values <u>given the data</u>.

Where it is known from other measurements that a DNA fraction is homogeneous or where a single copy fraction has been purified the least squares analysis provides an excellent means for evaluating its rate constant and permits a more accurate calculation of its complexity. In general it is not known that components are homogeneous and often the

values of parameters derived from the least squares solutions may be averages of a set of unresolved components. In such a case the solution makes an excellent model of the set of repetitive sequences which may be used in a variety of calculations even though the true individual components are not known. Where an important issue rests on the potential heterogeneity of a component other tools must be used. For example fractionation of double from single standed DNA could be carried out at the midpoint of its reassociation and the rate of reassociation of the two fractions carefully compared.

Another consideration is as important as questions of parameter meaning: solution uniqueness. If a parameter can change over a wide range without affecting the quality of the fit measured by the RMS or GOODNESS OF FIT, conclusions based on a specific parameter value are suspect. In many cases, this problem will be indicated by a high parameter standard deviation. More quantitative insight into the significance of a particular parameter value can be gained by plotting the RMS or GOODNESS OF FIT criterion as a function of the best fit using different fixed values for the parameter in question. For example, if the best fit of the data gives a rate constant of 0.05 for a slow component with a GOODNESS OF FIT of 0.02, other solutions should be found with the rate constant fixed at 0.15 and 0.015. If the variation in the GOODNESS OF FIT is less than 10% over this range of rate constants, no conclusions can be based on the 0.05 value which would be different for the other values. If the GOODNESS OF FIT changes by 50- to 100% with variations in the parameter value the value is probably uniquely determined by the data. While the method of fixing one parameter value and varying the others to find the best least squares

solution can be used for any parameter, it is particularly important when measuring rate constant parameter values. Data with two rate constants differing by a factor of 10-30 can usually be described by a very wide range of rate constants.

In summary, once a fully convergent least squares solution has been established for a set of data we must consider four issues of interpretation: 1. There may be systematic errors in the measurement such as DNA degradation or unknown fragment size. 2. The individual set of data may be atypical, particularly if only a few measurements are available and the standard deviations are then a poor measure of the possible error. Deviations in reassociation kinetic measurements are typically not due to random statistical variables such as sampling from a population or radioactive decay, but are more likely due to variations in assay procedures such as hydroxyapatite batch or temperature or volume of samples. This weakens the significance of the standard deviations. 3. The components may not represent true individual rate components but be averages of unresolved sets of components. 4. The least squares minimum in the region of the solution may be shallow. In that case, the set of solution parameter values may not uniquely fit the data; other parameter sets with substantially different values might provide equally valid interpretations of the data.

In many cases such problems can be shown to be of minor quantitative significance. Least squares solutions such as those generated by this program represent the best presently known approach to the interpretation of reassociation and hybridization kinetics.

References Cited

Bishop, J. O., Morton, J. G., Roshbash, M. and Richardson, M. (1974) "Three abundance classes in HeLa cell messenger RNA" Nature 250:199-204

Britten, R. J. and Kohne, D. E. (1968) "Repeated sequences in DNA", Science 161:529-540

Britten, R. J., Graham, D. E. and Neufeld, B. R. (1974) "Analysis of repeating DNA sequences by hybridization" in Methods in Enzymology L. Grossman and K. Moldave, eds. Vol. 29 Part E pp. 363-418

Britten, R. J. and Davidson, E. H., (1976) "Studies on nucleic acid reassociation kinetics: Empirical equations describing DNA reassociation" Proc. Nat. Acad. Sci. US 73:415-419

Davidson, E. H. and Britten, R. J., (1973) "Organization, transcription and regulation in the animal genome" Quart. Rev. Biol. 48:565-613

Davidson, E. H., Galau, G. A., Angerer, R. C. and Britten, R. J., (1975) "Comparative aspects of DNA organization in metazoa" Chromosoma 51:253-259

Davidson, E. H., Hough, B. R., Klein, W. H. and Britten, R. J. (1975) "Structural genes adjacent to interspersed repetitive DNA sequences" Cell 4:217-238

Galau, G. A., Britten, R. J. and Davidson, E. H., (1974) "A measurement of the sequence complexity of polysomal messenger RNA in sea urchin embryos" Cell 2:9-20

Galau, G. A., Klein, W. H., Davis, M. M., Wold, B. J. Britten, R. J. and Davidson, E. H. (1976) "Structural gene sets active in embryos and adult tissues of the sea urchin" Cell 7:487-505

Laird, C. D. (1971) "Chromatid structure: Relationships between DNA content and nucleotide sequence complexity" Chromosoma 32:378-406

Marquardt, D. W. (1963) "An algorithm for least squares estimation of non-linear parameters" J. Soc. Indust. Appl. Math. 11:431-441

Martin, R. S., Peters, G., and Williamson, J. H. (1971) "Symmetric decomposition of a positive definate matrix" in Linear Algebra by J. H. Wilkinson and C. Reinsch Vol. II of HANDBOOK FOR AUTOMATIC COMPUTATION F. L. Bauer, A. S. Householder, F. W. J. Olver, H. Rutishauser, K. Samelson and E. Statel, ed. New York: Springer-Verlag pp. 9-30

Morrow, J. (1974) Ph. D. Thesis, Stanford University

Ryffel, G. U. and McCarthy, B. J. (1975) "Complexity of cytoplasmic RNA in different mouse tissues measured by hybridization of polyadenylated RNA to complementary DNA" Biochemistry 14:1379-1384

Smith, M. J., Britten, R. J. and Davidson, E. H. (1975) "Studies on nucleic acid reassociation kinetics: Reactivity of single stranded DNA tails in DNA-DNA renaturation" Proc. Nat. Acad. Sci. US 72:4805-4809

APPENDIX I

This is a sample run of the program with user responses underlined

.RU NNNBAT

```
 1START,2ADD,3GUESS,5PLOT,6CPLOT,8LPLOT,9STOP

1                                   start by getting data
TYPE RED DATA FILENAME

RTDNA                       name of file on disk
106 PT RAT COT CURVE
TO LIST DATA TY 1

__                                  do not list data


 1START,2ADD,3GUESS,5PLOT,6CPLOT,8LPLOT,9STOP

3                                   do a fit
FUNCTION (-1HELP), ITERATIONS, RMS QUIT, DELMX QUIT, (NIT)

-1                                  check the names of the functions
0SAME,1FINGER,2WHATOR,3NUFORM,4EXCESS,5WHTCMP

FUNCTION (-1HELP), ITERATIONS, RMS QUIT, DELMX QUIT, (NIT)

1,10,.0001,.05
FINAL,FRACT(1),K(1),FRACT(2),K(2),FRACT(3),K(3)

.05 .17 2. .08 .01 .65 .0005    initial parameter guesses
TYPE INDEX OF PARAMETERS TO BE FIXED.

__                                  no fixed parameters

ORIGINAL RMS=    0.0447680
GOODNESS OF FIT    0.0463237


     RMS      DELMX      PARAMETERS

 0.0304963 1.7717584
 0.393E-01 0.171E+00 0.722E+00 0.979E-01 0.985E-02 0.602E+00 0.455E-03

 0.0298000 0.2591739
 0.383E-01 0.174E+00 0.974E+00 0.100E+00 0.867E-02 0.598E+00 0.445E-03

 0.0297149 0.0705356
```

```
 0.374E-01 0.176E+00 0.105E+01 0.103E+00 0.817E-02 0.594E+00 0.438E-03

        iterations stop because RMS is
        changing by less than RMS QUIT=0.0001

        covariance matrix

I       1           2           3           4           5           6           7

 1 0.100E+01
 2-0.251E+00 0.100E+01
 3 0.175E+00-0.612E+00 0.100E+01
 4-0.699E+00 0.312E+00-0.180E+00 0.100E+01
 5 0.559E+00-0.715E+00 0.555E+00-0.825E+00 0.100E+01
 6 0.353E+00-0.526E+00 0.385E+00-0.876E+00 0.901E+00 0.100E+01
 7 0.861E+00-0.406E+00 0.289E+00-0.941E+00 0.820E+00 0.759E+00 0.100E+01


106 PT RAT COT CURVE


RMS=     0.0297149 GOODNESS OF FIT    0.0307474
    106 POINTS       99 DEG OF FREEDOM


..P(1)......P(2)......P(3)......P(4)......P(5)......P(6)......P(7)...

        parameters are printed in the same order
        they were input.  FINAL,FRACT(1),K(1),...
        the first line is the parameter values

 0.374E-01 0.176E+00 0.105E+01 0.103E+00 0.817E-02 0.594E+00 0.438E-03
 0.325E-01 0.198E-01 0.451E+00 0.695E-01 0.116E-01 0.609E-01 0.140E-03

        the second line is the parameter errors

 1START,2ADD,3GUESS,5PLOT,6CPLOT,8LPLOT,9STOP

3                                   Do another fit to illustrate
                                    OSAME function and a local
                                    minimum without convergence
FUNCTION (-1HELP), ITERATIONS, RMS QUIT, DELMX QUIT, (NIT)

0,10,0.,.01                         OSAME, the fit continues from
                                    where it left off
ORIGINAL RMS=    0.0297149
GOODNESS OF FIT    0.0307474


     RMS      DELMX      PARAMETERS

 0.0296630 0.0514820
 0.366E-01 0.177E+00 0.106E+01 0.107E+00 0.777E-02 0.591E+00 0.431E-03

 0.0296151 0.0407157
```

```
 0.358E-01 0.177E+00 0.106E+01 0.110E+00 0.746E-02 0.589E+00 0.425E-03

 0.0295696 0.0350217            Note the change in DELMX after
 0.349E-01 0.177E+00 0.106E+01 0.114E+00 0.721E-02 0.586E+00 0.419E-03

 0.0295148 0.6947887            this point. A local minimum was found,
 0.264E-01 0.179E+00 0.103E+01 0.143E+00 0.425E-02 0.563E+00 0.367E-03

 0.0290103 0.4464171            going on caused a shift to a path
 0.182E-01 0.175E+00 0.112E+01 0.166E+00 0.531E-02 0.553E+00 0.331E-03

 0.0288781 1.4516340            leading to a negative parameter
 0.744E-02 0.177E+00 0.106E+01 0.189E+00 0.382E-02 0.538E+00 0.295E-03

 0.0286363 6.6211100
-0.132E-02 0.175E+00 0.112E+01 0.208E+00 0.402E-02 0.530E+00 0.268E-03

 0.0285198 0.8757134
-0.106E-01 0.176E+00 0.109E+01 0.225E+00 0.351E-02 0.521E+00 0.244E-03

 0.0284246 0.4295531
-0.187E-01 0.176E+00 0.110E+01 0.240E+00 0.337E-02 0.515E+00 0.225E-03

 0.0283610 0.2825794            now the fit will not converge
-0.260E-01 0.176E+00 0.110E+01 0.252E+00 0.317E-02 0.510E+00 0.209E-03

I      1          2          3          4          5          6          7

 1 0.100E+01
 2-0.177E+00 0.100E+01
 3 0.132E+00-0.380E+00 0.100E+01
 4-0.722E+00 0.345E+00-0.251E+00 0.100E+01
 5 0.535E+00-0.607E+00 0.490E+00-0.885E+00 0.100E+01
 6 0.967E-01-0.455E+00 0.358E+00-0.747E+00 0.828E+00 0.100E+01
 7 0.900E+00-0.305E+00 0.233E+00-0.937E+00 0.774E+00 0.498E+00 0.100E+01


106 PT RAT COT CURVE

RMS=    0.0283610 GOODNESS OF FIT    0.0293465
    106 POINTS      99 DEG OF FREEDOM

..P(1)......P(2)......P(3)......P(4)......P(5)......P(6)......P(7)...

-0.260E-01 0.176E+00 0.110E+01 0.252E+00 0.317E-02 0.510E+00 0.209E-03
 0.565E-01 0.135E-01 0.383E+00 0.829E-01 0.176E-02 0.635E-01 0.101E-03

 1START,2ADD,3GUESS,5PLOT,6CPLOT,8LPLOT,9STOP

3_                              a fit to show parameter fixing
FUNCTION (-1HELP), ITERATIONS, RMS QUIT, DELMX QUIT, (NIT)

0,15,0. 0.
FINAL,FRACT(1),K(1),FRACT(2),K(2),FRACT(3),K(3)
```

.05 .17 2. .08 .01 .65 .00035 k(3) is set to rate calculated from genome size

TYPE INDEX OF PARAMETERS TO BE FIXED.

7 hold parameter 7 constant. note that with this constraint the fit converges

ORIGINAL RMS= 0.0608586
GOODNESS OF FIT 0.0626578

RMS DELMX PARAMETERS

0.0438545 7.8186761
0.997E-02 0.176E+00 0.496E+00 0.134E+00 0.113E-02 0.589E+00 0.350E-03

0.0340088 0.6862643
0.318E-01 0.186E+00 0.772E+00 0.108E+00 0.315E-02 0.583E+00 0.350E-03

0.0293703 0.5452864
0.235E-01 0.170E+00 0.111E+01 0.149E+00 0.692E-02 0.570E+00 0.350E-03

0.0292464 0.6662786
0.228E-01 0.178E+00 0.105E+01 0.158E+00 0.415E-02 0.553E+00 0.350E-03

0.0290227 0.2421386
0.301E-01 0.175E+00 0.112E+01 0.171E+00 0.468E-02 0.536E+00 0.350E-03

0.0290199 0.1418406
0.334E-01 0.177E+00 0.107E+01 0.176E+00 0.410E-02 0.525E+00 0.350E-03

0.0290172 0.0786546
0.320E-01 0.176E+00 0.111E+01 0.174E+00 0.445E-02 0.530E+00 0.350E-03

0.0290166 0.0502960
0.330E-01 0.177E+00 0.109E+01 0.176E+00 0.424E-02 0.526E+00 0.350E-03

0.0290163 0.0301044
0.325E-01 0.176E+00 0.110E+01 0.175E+00 0.437E-02 0.529E+00 0.350E-03

0.0290163 0.0187524
0.328E-01 0.177E+00 0.109E+01 0.175E+00 0.429E-02 0.527E+00 0.350E-03

0.0290162 0.0113841
0.326E-01 0.176E+00 0.110E+01 0.175E+00 0.434E-02 0.528E+00 0.350E-03

0.0290162 0.0070176
0.327E-01 0.176E+00 0.109E+01 0.175E+00 0.431E-02 0.528E+00 0.350E-03

0.0290162 0.0042845
0.327E-01 0.176E+00 0.110E+01 0.175E+00 0.433E-02 0.528E+00 0.350E-03

0.0290162 0.0026309
0.327E-01 0.176E+00 0.109E+01 0.175E+00 0.431E-02 0.528E+00 0.350E-03

0.0290162 0.0016100

```
 0.327E-01 0.176E+00 0.110E+01 0.175E+00 0.432E-02 0.528E+00 0.350E-03

I      1          2          3          4          5          6
 1 0.100E+01
 2 0.216E+00 0.100E+01
 3-0.163E+00-0.413E+00 0.100E+01
 4 0.773E+00 0.131E+00-0.468E-01 0.100E+01
 5-0.547E+00-0.654E+00 0.529E+00-0.664E+00 0.100E+01
 6-0.900E+00-0.379E+00 0.291E+00-0.918E+00 0.807E+00 0.100E+01


106 PT RAT COT CURVE


RMS=     0.0290162 GOODNESS OF FIT    0.0298740
    106 POINTS      100 DEG OF FREEDOM


..P(1)......P(2)......P(3)......P(4)......P(5)......P(6)......P(7)...

 0.327E-01 0.176E+00 0.110E+01 0.175E+00 0.432E-02 0.528E+00 0.350E-03
 0.192E-01 0.142E-01 0.396E+00 0.300E-01 0.223E-02 0.509E-01 0.000E+00
               the covariance matrix does not include
        parameter 7 and the parameter 7 error is 0.000 because
        it is fixed


 1START,2ADD,3GUESS,5PLOT,6CPLOT,8LPLOT,9STOP

3                                   error estimates without iteration
FUNCTION (-1HELP), ITERATIONS, RMS QUIT, DELMX QUIT, (NIT)

1,0                                 0 iterations gives error estimates
FINAL,FRACT(1),K(1),FRACT(2),K(2),FRACT(3),K(3)

.0327 .176 1.10 .175 .00432 .528 .00035
TYPE INDEX OF PARAMETERS TO BE FIXED.

__                          check the error without the parameter fixed



I      1          2          3          4          5          6          7

 1 0.100E+01
 2-0.201E+00 0.100E+01
 3 0.145E+00-0.460E+00 0.100E+01
 4-0.735E+00 0.361E+00-0.254E+00 0.100E+01
 5 0.579E+00-0.627E+00 0.496E+00-0.900E+00 0.100E+01
 6 0.465E+00-0.484E+00 0.368E+00-0.933E+00 0.937E+00 0.100E+01
 7 0.877E+00-0.341E+00 0.252E+00-0.959E+00 0.828E+00 0.816E+00 0.100E+01


106 PT RAT COT CURVE
```

```
RMS=    0.0290170 GOODNESS OF FIT    0.0300253
    106 POINTS      99 DEG OF FREEDOM


..P(1)......P(2)......P(3)......P(4)......P(5)......P(6)......P(7)...

 0.327E-01 0.176E+00 0.110E+01 0.175E+00 0.432E-02 0.528E+00 0.350E-03
 0.401E-01 0.152E-01 0.414E+00 0.106E+00 0.399E-02 0.883E-01 0.160E-03

 1START,2ADD,3GUESS,5PLOT,6CPLOT,8LPLOT,9STOP

9
STOP
```

Appendix II

The program: Structure and Method

Appendix III provides an annotated listing of the program. The main program is divided into four blocks corresponding to input (1START,2GUESS), least squares fitting (3GUESS), plot output (5PLOT,8LPLOT) and data output for future plotting (6CPLOT). The input section scans the input until a 3.0 is found in the C/Co value and then sorts the data using the subroutine SORTI. Plotting is done with the subroutine NPLMLT. To plot the data on the line printer the unit number for the output statements is changed and the program goes to the 5PLOT section.

In the 3GUESS option, the input of the FUNCTION line and the parameter guesses are handled by the PARMIO subroutine. The number of parameters to be fit is counted from the input data by COUNT, the index of parameters to be fixed is read in and an array of parameters to be varied is generated by MAP. Once the parameters are specified, the main fitting routine, REFIN, is called with one argument, DIFSUB, used to specify which of the the subroutines FINDIF, WHTDIF, FRMDIF, EXCDIF or WHADIF should be used to provide the actual derivatives and function values for the fitting routine.

The algorithm used in REFIN was developed by Marquardt (1963) for non-linear least squares problems. The algorithm first attempts to find improved parameter values using Taylor's expansion around the region of the old parameter estimates. If the new parameters improve the fit they

are used for the subsequent iteration, if not, the parameter change vector is rotated in the direction of the decreasing error gradient until a new parameter set better fits the data. Two subroutines, (SYMDET and SYMSOL), manipulate the symmetric partial derivative matrices used to solve for the parameter changes. These programs were adapted from ALGOL versions published by Martin et. al. (1971). When the iterations terminate, the partial derivative matrix is again calculated and inverted by SYMINV to provide the error estimates and correlation matrix. WCOVAR writes out the correlation matrix and control is returned to the main program, where the parameters and error estimates are printed.

REFIN is particularly efficient in storage and manipulation of the partial derivative matrices. Only one new matrix with dimensions (number of variable parameters)^2 is stored. This results in significant savings of space over methods which store two square partial derivative matrices, one for the matrix and one for its inverse, and also a matrix storing the partial derivative for the jth parameter at the ith data point with dimensions (number of parameters * number of data points). This latter matrix grows with the number of data points and thus with the quality of the fit, requiring more storage for better data. While this is a minor consideration for nucleic acid experiments where more than 7 parameters and 50 points are rarely fit, it becomes important for the analysis of other data (particularly protein bands on polyacrylamide gels fit by Gaussian functions) where as many as 100 parameters may be fit to 1000 data points.

Parameter fixing uses an array, IMAP, to map the indices of parameters to be varied into the partial derivative matrices. Partial derivatives at each data point are calculated for all parameters, whether they are varied or not, but when the partial derivative matrices are built up, only the parameters with indices in the array IMAP are used. This provides a versatile method for holding any number or combination of parameters constant during the fit. Parameter fixing is particularly useful in careful analyses of the significance of particular sets of parameter values.

Fitting a number of functions decreases the computational efficiency of this program by requiring a subroutine call for each data point during the generation of the partial derivative matrix. A more efficient program fitting one function can be written putting the loop from the appropriate DIFSUB subroutine into REFIN.

Analytic derivatives are used for four of the five functions fit. Empirical derivatives used for the NUFORM function are calculated by successive evaluation of the function at one parameter value and a value 1% different. It is important to note that this partial differentiating routine can be used to fit any other function as well, simply by substituting the new function for the subroutine FORM.

This program, and the REFIN fitting subroutine in particular, can provide a versatile tool whenever non-linear least squares analysis is required. The algorithm used is efficient, the internal storage required is compact and the ability to fix any combination of parameters provides unique insight into the physical meaning of the solution.

APPENDIX III

```
C
C               (C) COPYRIGHT 1977      WILLIAM PEARSON
C
C     THIS VERSION SUPPORTS FIXED PARAMETERS, ERROR ESTIMATES
C     AND ERRORS FROM BATCH. IT HAS BEEN TESTED FOR CORRECT ERROR
C     ESTIMATES AND PARTIAL DERIVATIVES FOR FINGER,WHATOR,
C     WHTCMP AND EXCESS.
C
C     SUBROUTINES CALLED
C
C     REFIN,PARMIO,MAP,RRJE,COUNT,PLMLT,SORTI,FORM,WCOVAR
C
C     PROBLEMS WITH THIS PROGRAM MAY BE ANSWERED BY
C             BILL PEARSON
C             KERCHOFF MARINE LAB
C             101 DAHLIA
C             CORONA DEL MAR, CA. 92625
C
C             PH. (714) 675-2159
C
C     VARIABLE USAGE: ARRAYS
C
C             COT(200)        COT VALUES
C             YY(200)         FRACTION SINGLE STRANDED
C             NMAX=200        MAXIMUM NUMBER OF DATA POINTS
C
C             P(11)           PARAMETER VALUES
C             D(11)           PARAMETER ERRORS, PARAMETER CHANGE
C                             VECTOR FOR 4 FINGR
C             NPMAX=11        MAXIMUM NUMBER OF PARAMETERS
C
C             IA(11)          ARRAY OF INDICES FOR PARAMETERS HELD
C                             CONSTANT
C             IMAP(11)        MAPPING OF INDICES OF PARAMETERS ALLOWED
C                             TO CHANGE
C             ROUT(18)        ARRAY OF VALUES TO BE PLOTTED
C
C
C                     COMMON
C
C             /IO/            COMMON BLOCK FOR INPUT/OUTPUT  DEVICE
C                             NUMBERS
C
C             LRD             INPUT DEVICE; 5 FOR TTY TIMEHARING,
C                             PROBABLY ALSO 5 FOR CDR IN BATCH MODE
C             LWD             OUTPUT DEVICE; 5 FOR TTY TIMESHARING,
C                             PROBABLY 6 FOR LPT IN BATCH
C             LSD             STORAGE DEVICE; 20 FOR DSK ON PDP-10
C                             DATA IS STORED IN DSK FILES.
C                             OPEN SETS UP DSK FILES FOR READS AND WRITES
```

```
C                              PROBABLY 5 IN A BATCH SYSTEM.  NO
C                              OUTPUT TO LSD IS NECESSARY UNLESS AN
C                              OPTION 6 SYSTEM IS SET UP
C
C              :EXTERNAL SUBROUTINES
C
C     THIS PROGRAM WILL FIT 5 DIFFERENT FUNCIONS ENCOUNTERED
C     ED IN DNA-DNA AND DNA-RNA HYBRIDIZATION EXPERIMENTS.  THE
C     DIFFERENT FUNCTIONS ARE SPECIFIED BY THE VALUE OF IWHAT=1..5
C     AND PARTIAL DERIVATIVES AND ERRORS ARE EVALUATED BY THE
C     EXTERNAL SUBROUTINES FINDIF,WHADIF,FRMDIF,EXCDIF,WHTDIF.
C
C     THE FUNCTIONS ARE:
C
C     IWHAT   EXTERNAL          FUNCTION
C
C       1      FINDIF           MULTICOMPONENT  2ND ORDER REASSOCIATION
C             (FINGER)          AS MEASURED ON HYDROXYLAPATITE.
C
C                               F(X,P)=P(1)+P(2I)/(1.P(2I+1)*X)
C                                 FOR I=1 TO NC COMPONENTS
C
C       2      WHADIF           FUNCTION USED TO STUDY KINETCS OF
C             (WHATOR)          DNA REASSOCIATION MEASURED BY NUCLE-
C                               ASE DIGESTION.
C
C                               F(X,P)=P(1)+P(2)/(1.+P(3)*X)**P(4)
C
C       3      FRMDIF           FUNCTION WHICH DESCRIBES THE HYBRIDI-
C             (NUFORM)          ZATION OF UP TO 2 DRIVEN COMPONENTS
C                               WITH RATES P(3),P(5) DRIVEN BY A VAST
C                               EXCESS OF A COMPONENTS HYBRIDIZING WITH
C                               RATE P(6).   P(2,4) ARE THE FRACTIONS OF
C                               THE COMPONENTS DRIVEN. P(7) IS THE
C                               FACTOR ANALOGOUS TO P(4) IN WHATOR.
C                               P(1)=THE FINAL AMOUNT UNREACTED.
C
C                               F(X,P)=P(1)+P(2I)*EXP(P(2I+1)*(
C                     1.-(1+P(6)*X)**(1-P(7)))/(P(6)*(1.-P(7))))
C
C       4      EXCDIF           1ST ORDER REASSOCIATION (RNA EXCESS)
C             (EXCESS)
C                               F(X,P)=P(1)+P(2I)*EXP(-P(2I+1)*X)
C
C       5      WHTDIF           MULTICOMPONENT REASSOCIATION MEASURED BY
C             (WHTCMP)          NUCLEASE
C                               F(X,P)=P(1)+P(2I)/(1.+P(2I+1)*X)**(0.44)
C
C     F(X,P) IS EACH CASE IS THE FRACTION SINGLE STRANDED AT COT X.
C     THIS PROGRAM CAN FIT OR PLOT UP TO 5 COMPONENTS FOR FINGER,
C     EXCESS OR WHTCOMP AS THE NUMBER OF PARAMETERS IN THOSE CASES
C     IS NP=2*NC+1 AND NPMAX=11.
```

```
C
C
C     NOTE, STATEMENTS FLAGGED:
*--------------------
C     ARE SUBROUTINES OR OPTIONS POSSIBLY SPECIFIC  FOR THE PDP-10
C     AND MAY NOT WORK ON OTHER SYSTEMS
C
C
      DIMENSION COT(200),YY(200),P(11),D(11),IMAP(11),IA(11),ROUT(18)
      DOUBLE PRECISION CNAME,TITLE
      COMMON /IO/LRD,LWD,LSD
      EXTERNAL FINDIF,WHADIF,FRMDIF,EXCDIF,WHTDIF
      DATA NMAX,NPMAX/200,11/
C
C     COMCON IS THE CONSTANT RELATING
C     RATE CONSTANT TO NUCLEOTIDES COMPLEXITY
C
      DATA COMCON/1015000./
C
      LRD=5
      LWD=5
      LSD=20
C
C     LWDSAV IS USED TO SAVE THE WRITE DEVICE WHEN IS IS
C     CHANGED IS OPTION 8.  THIS  WAY ONLY THE ABOVE THREE STATEMENTS
C     NEED BE CHANGED TO CHANGE IO DEVICES
C
      LWDSAV=LWD
C
C     LPTD=3 IS TO USE THE LINE PRINTER IN OPTION 8
C
      LPTD=3
      IWHAT=1
C
C     MAIN SWITCH POINT.  AFTER ANY OPTION IS DONE, RETURN
C     HERE TO CHOOSE NEW OPTION
C
   10 WRITE(LWD,20)
   20 FORMAT('0 1START,2ADD,3GUESS,5PLOT,6CPLOT,8LPLOT,9STOP'/)
      READ(LRD,30) ICHO
*--------------------
C  30 FORMAT(I2)
   30 FORMAT(I)
      GOTO (100,105,200,10,500,600,10,800,900),ICHO
      GOTO 10
C
C     INPUT OPTION
C             GET A DATA FILE NAME
C             OPEN FILE FOR READING
C             READ TITLE
C             READ FRACT SS, COT AND COUNT NUMBER OF
C               POINTS UNTIL FRACT SS=3.
```

```
C
C     THE FORM OF THE INPUT DATA FILE SHOULD BE
C
C     FIRST CARD:      ANY 54 CHAR TITLE, 1ST CHAR BLANK
C     SECOND CARD:     FRACT. SINGLE STRANDED,COT.
C     NEXT CARDS:      SAME
C           ..         ..
C     LAST CARD:       3.0
C
  100 KEND=0
  105 WRITE(LWD,110)
  110 FORMAT(' TYPE RED DATA FILENAME'/)
      READ(LRD,115) TITLE
  115 FORMAT(A10)
*--------------------
      OPEN(UNIT=LSD,ACCESS='SEQIN',FILE=TITLE)
      READ(LSD,120)
  120 FORMAT(54H------------------------------------------------------)
      WRITE(LWD,120)
  125 KEND=KEND+1
      READ(LSD,130) YY(KEND),COT(KEND)
*--------------------
C 130 FORMAT(1X,F7.5,E11.4)
  130 FORMAT(2F)
      IF (YY(KEND).GE.3.0) GOTO 135
      IF (KEND.GT.NMAX) GOTO 160
      GOTO 125
  135 KEND=KEND-1
C
C     SORT DATA IN INCREASING ORDER BY COT.
C      NECESSARY FOR PLOTTING
C
      CALL SORTI(COT,KEND,YY)
      WRITE(LWD,140)
  140 FORMAT(' TO LIST DATA TY 1'/)
      READ(LRD,145) ICHO
*--------------------
C 145 FORMAT(I2)
  145 FORMAT(I)
      IF (ICHO.NE.1) GOTO 10
      WRITE(LWD,155)
  155 FORMAT(8X,'COT',2X,'FRACT SS'//)
      WRITE(LWD,150) (COT(I),YY(I),I=1,KEND)
  150 FORMAT(5X,F10.3,F6.3)
      GOTO 10
C
C     TOO MUCH DATA, ONLY READ IN FIRST NMAX POINTS
C
  160 WRITE(LWD,165)
  165 FORMAT(' MAX NUMBER OF POINTS EXCEEDED'/' SOME DATA LOST'/)
      GOTO 135
C
```

```
C     GUESS OPTION FOR PARTIAL DERIVATIVE LEAST SQUARES FIT
C     INPUT IWHAT, PARAMETERS AND CALCULATE NP(NO. OF PARAM).
C     AND NC(NO. OF COMPONENTS).
C
  200 CONTINUE
  210 CALL PARMIO(P,NP,NC,NPMAX,IWHAT,NINT,RMSQU,DLMXQU,NIT,&228)
      WRITE(LWD,220)
  220 FORMAT(' TYPE INDEX OF PARAMETERS TO BE FIXED.'/)
      READ(LRD,225) (IA(I),I=1,NPMAX)
*-------------------
C 225 FORMAT(11I2)
  225 FORMAT(11I)
C
C     MAP PARAMETERS TO BE VARIED IN IMAP.
C
      CALL MAP(IA,NP,IMAP,NDIFF)
  228 CONTINUE
C
C     CALL LEAST SQUARES FITTING ROUTINE USING APPROPRIATE
C     PARTIAL DERIVATIVE SUBROUTINE.
C
      GOTO (255,260,265,270,275),IWHAT
  255 CALL REFIN(COT,YY,KEND,P,D,NP,NC,IMAP,NDIFF,NIT,0.,FINDIF,DIFFSQ,
     1NINT,RMSQU,DLMXQU)
      GOTO 290
  260 CALL REFIN(COT,YY,KEND,P,D,NP,NC,IMAP,NDIFF,NIT,0.,WHADIF,DIFFSQ,
     1NINT,RMSQU,DLMXQU)
      GOTO 290
  265 CALL REFIN(COT,YY,KEND,P,D,NP,NC,IMAP,NDIFF,NIT,0.,FRMDIF,DIFFSQ,
     1NINT,RMSQU,DLMXQU)
      GOTO 290
  270 CALL REFIN(COT,YY,KEND,P,D,NP,NC,IMAP,NDIFF,NIT,0.,EXCDIF,DIFFSQ,
     1NINT,RMSQU,DLMXQU)
      GOTO 290
  275 CALL REFIN(COT,YY,KEND,P,D,NP,NC,IMAP,NDIFF,NIT,0.,WHTDIF,DIFFSQ,
     1NINT,RMSQU,DLMXQU)
      GOTO 290
  290 CONTINUE
C
C     OUTPUT TITLE , PARAMETERS AND PARAMETER ERRORS.
C
      WRITE(LWD,120)
      IDENOM=KEND-NDIFF
      IF (IDENOM.LT.1) IDENOM=KEND
      RMS=SQRT(DIFFSQ/KEND)
      CHISQ=SQRT(DIFFSQ/IDENOM)
      WRITE(LWD,235) RMS,CHISQ,KEND,IDENOM
  235 FORMAT('0RMS=',F13.7,' GOODNESS OF FIT',F13.7/5X,I3,' POINTS',5X,
     1I3,' DEG OF FREEDOM'/)
      WRITE(LWD,230)
  230 FORMAT('0...P(1).....P(2).....P(3).....P(4).....P(5).....P(6)
     1.....P(7)...'/)
```

```
      WRITE(LWD,240) (P(I),I=1,NP)
  240 FORMAT(1X,10E10.3,/)
      WRITE(LWD,240) (D(I),I=1,NP)
      GOTO 10
C
C     OPTION 5 PLOT.  PLOTS DATA AND APPROPRIATE FUNCTION ON TTY
C     WITH A LOG SCALE.
C             LPST=8 IS FIRST POSITION FOR DATA
C             LPSP=18 IS LAST POSITION FOR DATA
C             II IS BEGINNING X VALUE FOR PLOT
C             IEND IS EXPONENT OF LAST POINT PLOTTED.
C             K IS CURRENT DATA POINT TO BE PLOTTED
C
  500 CALL PLMLT(X,ROUT,L,1,100.,0.)
      LPST=8
      LPSP=18
      II=12.*ALOG10(COT(1))
      IEND=(ALOG10(COT(KEND))-ALOG10(COT(1)))*12.+24.
      K=1
C
C     ZERO ARRAY OF VALUES TO BE PLOTTED
C
      DO 515 L=1,LPSP
  515 ROUT(L)=0.
C
C     FOR EACH PLOT INCREMENT ALONG THE X AXIS
C
      DO 510 J=1,IEND
      XL=FLOAT(II+J)/12.-1.
      X=10.**XL
      ROUT(1)=0.
      L=LPST-1
C
C     IF THERE ARE STILL DATA POINTS TO BE PLOTTED
C
      IF(K.GT.KEND) GOTO 530
C
C     CHECK TO SEE IF A DATA POINT SHOULD BE PLOTTED
C
  525 IF (ALOG10(COT(K)).GT.(XL+.04)) GOTO 530
C
C     STORE THE ERROR IN THE XVALUE OF THE POINT
C
      ROUT(1)=XL-ALOG10(COT(K))+.1
      L=L+1
C
C     STORE THE Y VALUES OF THE POINT AND INCREMENT THE
C     PLOT ARRAY INDEX(L) AND DATA ARRAY INDEX(K)
C
      ROUT(L)=YY(K)
      K=K+1
      IF (L.GE.LPSP) GOTO 530
```

```
      IF (K.LE.KEND) GOTO 525
C
C     CALCULATE THE COMPONENT AND FUNCTION VALUES AND
C     STORE IN PLOT ARRAY.
C
  530 CONTINUE
      GOTO (1540,1550,1560,1570,1580),IWHAT
C
C     FINGER
C
 1540 SUMC=P(1)
      DO 1542 I=1,NC
      JF=2*I
      JK=JF+1
      COMP=P(JF)/(1.0+X*P(JK))
      ROUT(I+2)=COMP
 1542 SUMC=SUMC+COMP
      ROUT(2)=SUMC
      GOTO 570
C
C     WHATOR
C
 1550 ROUT(3)=0.
      ROUT(5)=0.
      ROUT(4)=P(2)/(1.0+X*P(3))
      COMP=P(2)/(1.0+X*P(3))**P(4)
      ROUT(2)=COMP+P(1)
      GOTO 570
C
C     NUFORM
C
 1560 CALL FORM(P,X,ROUT(2))
      COMP=P(4)
      P(4)=1.0
      CALL FORM(P,X,ROUT(5))
      P(4)=COMP
      GOTO 570
C
C     EXCESS
C
 1570 SUMC=P(1)
      DO 1572 I=1,NC
      JF=2*I
      JK=JF+1
      XE=X*P(JK)
      COMP=0.
      IF (XE.LT.10.) COMP=P(JF)*EXP(-XE)
      ROUT(I+2)=COMP
 1572 SUMC=SUMC+COMP
      ROUT(2)=SUMC
      GOTO 570
C
```

```
C     WHTCMP
C
 1580 SUMC=P(1)
      DO 1582 I=1,NC
      JF=2*I
      JK=JF+1
      COMP=P(JF)/(1.+X*P(JK))**(0.44)
      ROUT(I+2)=COMP
 1582 SUMC=SUMC+COMP
      ROUT(2)=SUMC
      GOTO 570
C
C     PLOT DATA AND COMPONENTS IN ARRAY ROUT
C
  570 CALL PLMLT(X,ROUT,L,3,100.,0.)
  510 CONTINUE
C
C     PUT ON BOTTOM AXIS
C
      CALL PLMLT(X,ROUT,L,2,100.,0.)
      WRITE(LWD,572)
  572 FORMAT(1H0)
      WRITE(LWD,120)
      WRITE(LWD,235) RMS,CHISQ
      WRITE(LWD,575)
  575 FORMAT('0 COMPONENT...FRACTION...COMPLEXITY.....K.....
     1KPURE'/)
C
C     OUTPUT FINAL COMPLEXITIES
C
      DO 580 J=1,NC
      L=2*J
      Q=COMCON*P(L)/P(L+1)
      QQ=P(L+1)/P(L)
      WRITE(LWD,585) J,P(L),Q,P(L+1),QQ
  585 FORMAT(1H0,I8,F10.4,3E15.4)
  580 CONTINUE
      WRITE(LWD,590) P(1)
  590 FORMAT('0FINAL=',F7.4)
      LWD=LWDSAV
      GOTO 10
C
C     THIS SECTION WRITES DATA FILES FOR PLOTTING BY OTHER
C     PROGRAMS SUCH AS A CALCOMP PLOTTER PROGRAM.
C     IT MAY BE REMOVED ON OTHER SYSTEMS
C
  600 WRITE(LWD,610)
  610 FORMAT(' GIVE CALCOMP PLOT DATA FILE NAME'/)
      READ(LRD,615) CNAME
  615 FORMAT(A10)
*-------------------
      OPEN(UNIT=LSD,ACCESS='SEQOU',FILE=CNAME)
```

```
      WRITE(LWD,620)
  620 FORMAT(' TYPE TITLE FOR PLOT'/)
      READ(LRD,625)
  625 FORMAT(54H----------------------------------------------------)
      WRITE(LSD,625)
      WRITE(LSD,630) KEND,NP,IWHAT
  630 FORMAT(3I6)
      WRITE(LSD,633) (P(I),I=1,NP)
  633 FORMAT(5E16.9)
      WRITE(LSD,635) (COT(I),YY(I),I=1,KEND)
  635 FORMAT(5(E11.4,F5.3))
      END FILE LSD
      GOTO 10
C
C     THIS SECTION CHANGES LWD TEMPORARILY SO THAT THE PLOT OF
C     THE DATA IS WRITEEN OUT TO THE LINE PRINTER INSTEAD OF
C     THE TTY
C
  800 LWDSAV=LWD
      LWD=LPTD
      WRITE(LWD,810)
  810 FORMAT(1H1)
      WRITE(LWD,820) TITLE
  820 FORMAT('  FILE NAME= ',A10,//)
      WRITE(LWD,120)
      WRITE(LWD,230)
      WRITE(LWD,240) (P(I),I=1,NP)
      WRITE(LWD,240) (D(I),I=1,NP)
      WRITE(LWD,572)
      GOTO 500
  900 CONTINUE
      STOP
      END
C     REFBAT.F4       VERSION 306     20 AUG 1976
C
C     THIS VERSION 305 USES A BATCH METHOD AND SCALES THE
C     PARTIAL DERIVATIVE MATRIX AND USES THE COSINE
C     CORRECTION DONE CORRECTLY
C
C
C     THIS VERSION SUPPORTS FIXED PARAMETERS, CALCULATES
C     PARAMETER ERRORS CORRECTLY, AND ALLOWS FOR
C     TYPING OF A PARAMETER CORRELATION MATRIX.
C
C     THIS IS A GENERAL LEAST SQUARES FITTING PROGRAM USING A
C     MODIFIED DAMP LEAST SQUARES ALGORITHM DESIGNED BY MARQUARDT.
C     J. SOC. IND. APPL. MATH. (1965) 11:431-441
C
C     ARGUMENTS:
C
C             XP(1)   ARRAY OF DATA X VALUES
C             YP(1)   ARRAY OF DATA Y VALUES
```

```
C             N       NUMBER OF DATA VALUES
C
C             P(1)    PARAMETER VAUES
C             D(1)    PARAMETER ERRORS
C             NP      NUMBER OF PARAMETERSS
C
C             IMAP    INDEX MAPPING OF PARAMETERS TO BE VARIED
C             ND      NUMBER OF PARAMETERS VARIED
C
C             NIT     NUMBER OF CORRECTIVE ITERATIONS ALLOWED
C
C             FIN     RMS CHANGE CONSIDERED SIGNIFICANT,
C                     SHOULD BE 0.
C             SUBDIF  SUBROUTINE PROVIDING PARTIAL DERIVATIVES
C                     AND FUNCTION VALUES.
C             DIFFSQ  ERROR OF THE FIT
C
C             NINT    NUMBER OF ITERATIONS, IF LE 0 ERCAL=.TRUE.
C                     AND NO ITERATIONS
C
C             RMSQU   QUIT CRITERIA FOR RMS
C
C             DLMXQU  QUIT CRITERIA FOR DELTA MAX
C
C
C
C     THIS SUBROUTINE FORMS THE MATRICES PP AND PY AND THEN USES
C     THEM TO SOLVE THE SET OF EQUATIONS
C
C             PP*DELTA=PY
C
C     AND THEN CHANGES THE PARAMETER VALUES BY DELTA, I.E.
C
C             P(IMAP(I))=DELTA(I)
C
C     IMAP IS USED TO HOLD PARAMETERS CONSTANT.  A CONSTANT PARAMETER
C     INDEX IS NOT INCLUDED IN IMAP(I) SO THAT PARAMETER WILL NOT BE
C     CHANGED.  A MAPPING MUST BE USED FOR CORRECT CALCULATION
C     OF DELTA.
C
C     THE RMS OF THE NEW PARAMETER VALUES IS CALCULATED.
C     IF THE FIT IS IMPROVED THE ALGORITHM STARTS OVER. IF NOT,
C     THE DIAGONAL OF PP IS INCREASED AND NEW DELTAS CALCULATED.
C
C             PP(I,J)=PP(I,J)+PART(IMAP(I))*PART(IMAP(J))
C
C     SUMMED OVER ALL THE PARTIAL DERIVATIVES OF ALL THE DATA POINTS.
C
C             PY(I)=PY(I)+(F(X,P)-Y)*PART(IMAP(I)))
C
C     SUMMED OVER ALL THE DATA.
C
```

```
C     THE MATRIX PP IS SYMMETRIC SO ONLY THE UPPER RIGHT HAND
C     TRIANGLE IS STORED AND THE LOWER LEFT HAND
C     TRIANGLE AND THE ARRAY DIA IS USED TO STORE THE CHOLESKY
C     DECOMPOSITION OF THE MATRIX.  THE INVERSE OF THE MATRIX IS LATER
C     STORED THERE FOR THE ERROR CALCULATION.
C
C     THE MATRIX HANDLING ROUTINES WHICH DO THIS ARE SYMDET,SYMSOL
C     (WHICH SOLVES THE LINEAR EQUATIONS USING PY). AND SYMINV
C     WHICH INVERTS THE MATRIX.
C
      SUBROUTINE REFIN(XP,YP,N,P,D,NP,NC,IMAP,ND,NIT,FIN,SUBDIF,DIFFSQ,
     1NINT,RMSQU,DLMXQU)
      DIMENSION XP(1),YP(1),P(1),D(1),IMAP(1)
      DIMENSION PP(11,11),PY(11),DIA(11),PART(11),OP(11),DELTA(11)
      LOGICAL ERCAL
      COMMON/IO/LRD,LWD,LSD
C
C     ITNO HOLDS NUMBER OF REAL ITERATIONS
C     ITC HOLDS THE NUMBER OF CORRECTIVE ITERATIONS.
C     XLAM IS THE CORRECTION FACTOR
C     ERCAL IS A FLAG FOR DOING THE ERROR CALCULATION
C     GAMCRT IS THE CRITERION FOR USING THE COSINE CORRECTION METHOD
C
      GAMCRT=3.14159265/4.
      ERCAL= (NINT.LE.0)
      ITNO=0
      DENOM=N-ND
      IF (DENOM.GE.1) GOTO 14
      WRITE(LWD,12)
   12 FORMAT(' SOLUTION IS UNDERDETERMINED'/)
      DENOM=N
   14 IF (ERCAL) GOTO 40
C     CALCULATE ORIGINAL ERRORS
C
      ITC=0
      XLAM=.01
      DIFFSQ=0.
      DO 10 I=1,N
      CALL SUBDIF(XP(I),Y,P,PART,NP,NC,SUMC,2)
      DIFF=YP(I)-SUMC
   10 DIFFSQ=DIFFSQ+DIFF*DIFF
C
C     STORE IN RMSO FOR TERMINATION CALCULATION
C
      RMSO=SQRT(DIFFSQ/N)
      CHISQ=SQRT(DIFFSQ/DENOM)
      WRITE(LWD,15) RMSO,CHISQ
   15 FORMAT(' ORIGINAL RMS=',F13.7,/' GOODNESS OF FIT',F13.7/)
C     WRITE OUT FORMAT OF RMS AND PARAMS
C
      WRITE(LWD,20)
   20 FORMAT('0',5X,'RMS',5X,'DELMX',5X,'PARAMETERS')
```

```
C
C     MAIN LOOP: CALCULATE PY,PP MATRICES
C
   40 ITNO=ITNO+1
      DO 45 J=1,NP
   45 OP(J)=P(J)
      DIFFSQ=0.
      DO 50 I=1,ND
      PY(I)=0.
      DO 50 J=1,ND
   50 PP(I,J)=0.
      DO 100 I=1,N
      X=XP(I)
      Y=YP(I)
      CALL SUBDIF(X,Y,P,PART,NP,NC,SUMC,1)
      DIFF=Y-SUMC
      DO 80 J=1,ND
      PY(J)=PY(J)+DIFF*PART(IMAP(J))
      DO 80 K=J,ND
   80 PP(J,K)=PP(J,K)+PART(IMAP(J))*PART(IMAP(K))
  100 DIFFSQ=DIFFSQ+DIFF*DIFF
C
C     IF ERCAL THEN DO NOT CHANGE PARAMETERS
C
      IF(ERCAL) GOTO 245
C
C     SCALE THE MATRIX
C
      DO 105 I=1,ND
      D(I)=SQRT(PP(I,I))
      XTEM=0.
      IF (D(I).NE. 0.0) XTEM=PY(I)/D(I)
      PY(I)=XTEM
      DO 105 J=1,I
      XTEM=0.
      IF ((D(I).NE. 0.0).AND.(D(J).NE. 0.0)) XTEM=PP(J,I)/(D(I)*D(J))
      PP(J,I)=XTEM
  105 CONTINUE
C
C
      FLAM=0.
      XKGAM=1.0
C
C     CORRECTIVE LOOP
C
  110 FLAM=XLAM-FLAM
C
C     CORRECT PP MATRIX
C
      DO 120 I=1,ND
  120 PP(I,I)=FLAM+PP(I,I)
C
```

```
C     DO DECOMPOSITION AND SOLVE EQUATIONS FOR DELTA
C
      CALL SYMDET(PP,ND,DIA,DET,&210)
      CALL SYMSOL(PP,ND,DIA,PY,DELTA)
C
C     FORM NEW PARAMETERS
C
C
C     SCALE THE DELTAS FIRST
C
      DO 123 J=1,ND
      XTEM=0.0
      IF (D(J).NE.0.0) XTEM=DELTA(J)/D(J)
  123 DELTA(J)=XTEM
C
C
  125 DO 130 J=1,ND
  130 P(IMAP(J))=P(IMAP(J))+XKGAM*DELTA(J)
C
C        CALCULATE NEW CHI SQUARE
C
      DIFFN=0.
      DO 150 I=1,N
      CALL SUBDIF(XP(I),Y,P,PART,NP,NC,SUMC,2)
      DIFF=YP(I)-SUMC
  150 DIFFN=DIFFN+DIFF*DIFF
C
C     IF FIT IS BETTER, GOTO 160 ELSE GOTO 170
C
      IF (DIFFSQ-DIFFN-FIN) 170,160,160
C
C     FIT IS WORSE,  INCREMENT CORRECTION COUNTER
C     REPLASE  P(I) WITH OLD PARAMETERS
C     CHANGE CORRECTION FACTOR
C     LOOP AROUND
C
  170 ITC=ITC+1
      DO 173 J=1,NP
  173 P(J)=OP(J)
C
C
  175 IF (ITC.GT.NIT) GOTO 180
C
C     CALCULATE VALUES FOR COSINE CORRECTION
C
      GAMNUM=0.
      PYSQ=0.
      DELSQ=0.
      DO 178 I=1,ND
      PYTEM=PY(I)
      DELTEM=DELTA(I)
      GAMNUM=GAMNUM+PYTEM*DELTEM
```

```
      PYSQ=PYSQ+PYTEM*PYTEM
  178 DELSQ=DELSQ+DELTEM*DELTEM
      GAMNUM=GAMNUM/(SQRT(PYSQ)*SQRT(DELSQ))
      IF (GAMNUM .GT. 1.0) GAMNUM=1.0
      GAMNUM=ACOS(GAMNUM)
      IF (GAMNUM.LT.GAMCRT) GOTO 176
      XLAM=10.*XLAM
      GOTO 110
  176 CONTINUE
C     WRITE(LWD,179) GAMNUM
C 179 FORMAT('OUSING COSINE CRITERION, GAMMA=',E10.3)
      XKGAM=XKGAM/2.
      GOTO 125
  180 WRITE(LWD,185)
  185 FORMAT(' CORRECTIVE ITERATIONS EXCEEDED'//)
      GOTO 230
C
C     FIT IS BETTER.
C     RESETS CORRECTION COUNTER
C     DECREASE CORRECTION FACTOR
C     DISPLAY ERROR AND DECIDE TO GO ON
C
  160 ITC=0
      XLAM=XLAM/10.
C
C     CALCULATE THE MAXIMUM
C     RELATIVE CHANGE IN PARAMETERS FOR DISPLAY
C
      DELMX=0.
      DO 163 I=1,ND
      IF (P(IMAP(I)).NE.0.0) XTEM=ABS(DELTA(I)/P(IMAP(I)))
  163 IF (XTEM.GT.DELMX) DELMX=XTEM
      RMS=SQRT(DIFFN/N)
  164 WRITE(LWD,165) RMS,DELMX,(P(I),I=1,NP)
  165 FORMAT('0',2F10.7/1X,11E10.3)
C
C     TERMINATION TEST
C
      IF ((RMSO-RMS).LE.RMSQU) GOTO 230
      IF (DELMX.LE.DLMXQU) GOTO 230
      IF(ITNO.GE.NINT) GOTO 230
C
C     UPDATE OLD RMS
C
      RMSO=RMS
      GOTO 40
  210 WRITE(LWD,215)
  215 FORMAT(' EQUATIONS SINGULAR'/)
C
C     GO BACK ONE ITERATION AND QUIT
C
  220 DO 225 I=1,NP
```

```
  225 P(I)=OP(I)
C
C     DO THE ERROR CALCULATION FOR THE PARAMETERS
C
  230 ERCAL=.TRUE.
      GOTO 40
  245 DO 240 I=1,NP
  240 D(I)=0.0
      CALL SYMDET(PP,ND,DIA,DET,&200)
      CALL SYMINV(PP,ND,DIA)
      CHISQ=SQRT(DIFFSQ/DENOM)
      DO 250 I=1,ND
  250 D(IMAP(I))=SQRT(DIA(I))*CHISQ
C
C     REQUEST IF CORRELATION MATRIX SHOULD BE DISPLAYED.
C
      CALL WCOVAR(PP,DIA,IMAP,ND)
  200 RETURN
      END
C
C
C
C     SYMSUB.F4       VERSION 302 13-FEB-1976
C
C     THE FOLLOWING SUBROUTINES MANIPULATE SYMMETRIC MATRICES,
C     FINDING THE DETERMINANT, SOLVING LINEAR EQUATIONS,
C     AND INVERTING THE MATRIX.
C
C     THE FORTRAN PROGRAMS ARE MODIFICATIONS OF THE ALGOL
C     PROGRAMS PUBLISHED BY R.S. MARTIN, G. PETERS AND J.H.
C     WILKINSON 'SYMMETERIC DECOMPOSITION OF A POSITIVE DEFINATE
C     MATRIX" PP 9-25 IN "LINEAR ALGEBRA",  J.H. WILKINSON AND
C     C REINSCH ED.  VOL II OF HANDBOOK FOR AUTOMATIC COMPUTATION.
C     NEW YORK: SPRINGER-VERLAG 1971 439 PP.
C
C
CC    SYMDET FORM CHOLESKY DECOMPOSITION OF SYMMETERIC MATRIX
C     AND PLACE IN LOWER TRIANGLE AND ARRAY P.
C
C
      SUBROUTINE SYMDET(A,N,P,DET,*)
      DOUBLE PRECISION X,Y,Z,INPROD
      DIMENSION A(11,11),P(1)
      DET=1.
      DO 100 I=1,N
      DO 100 J=1,I
      X=A(J,I)
      IF (I.EQ.J) GOTO 10
      GOTO 50
   10 K=J-1
   15 IF (K.LT.1) GOTO 20
      Y=A(I,K)
```

```
      Z=Y*P(K)
      A(I,K)=Z
      X=X-Y*Z
      K=K-1
      GOTO 15
   20 DET=DET*X
      IF (X.EQ.0.) RETURN 1
      P(I)=1/X
      GOTO 100
   50 K=J-1
   60 IF (K.LT.1) GOTO 70
      INPROD=A(I,K)*A(J,K)
      X=X-INPROD
      K=K-1
      GOTO 60
   70 A(I,J)=X
  100 CONTINUE
      RETURN
      END
C
C
C
C     SYMSOL: SOLVE A SET OF LINEAR EQUATIONS  USING A LOWER
C     TRIANGULAR CHOLESKY DECOMPOSITION.
C
      SUBROUTINE SYMSOL(A,N,P,B,X)
      DOUBLE PRECISION Y,INPROD
      DIMENSION A(11,11),P(1),B(1),X(1)
      DO 50 I=1,N
      Y=B(I)
      K=I-1
   10 IF (K.LT.1) GOTO 20
      INPROD=A(I,K)*X(K)
      Y=Y-INPROD
      K=K-1
      GOTO 10
   20 X(I)=Y
   50 CONTINUE
      I=N
   55 IF (I.LT.1) GOTO 100
      Y=X(I)*P(I)
      K=I+1
   60 IF (K.GT.N) GOTO 70
      INPROD=A(K,I)*X(K)
      Y=Y-INPROD
      K=K+1
      GOTO 60
   70 X(I)=Y
      I=I-1
      GOTO 55
  100 CONTINUE
      RETURN
```

```
      END
C
C
C
C     SYMINV: INVERT A MATRIX USING A LOWER TRIANGULAR CHOLESKY
C     DECOMPOSITION MATRIX
C
      SUBROUTINE SYMINV(A,N,P)
      DIMENSION A(11,11),P(1)
      DOUBLE PRECISION X,Y,Z,INPROD
      DO 50 I=2,N
      DO 40 J=2,I
      J1=J-1
      X=-A(I,J1)
      K=I-1
   10 IF (K.LT.J) GOTO 20
      INPROD=A(I,K)*A(K,J1)
      X=X-INPROD
      K=K-1
      GOTO 10
   20 A(I,J1)=X
   40 CONTINUE
   50 CONTINUE
      DO 100 J=1,N
      DO 100 I=J,N
      IF (I.EQ.J) GOTO 70
      X=A(I,J)
      IF (I.EQ.N) GOTO 100
      I1=I+1
      DO 60 K=I1,N
      INPROD=A(K,I)*A(K,J)
   60 X=X+INPROD
      A(I,J)=X
      GOTO 100
   70 X=P(I)
      IF (I.EQ.N) GOTO 90
      I1=I+1
      DO 80 K=I1,N
      Y=A(K,J)
      Z=P(K)*Y
      A(K,J)=Z
      X=X+Y*Z
   80 CONTINUE
   90 P(I)=X
  100 CONTINUE
      RETURN
      END
C     ERPBAT.F4        VERSION 305     1 JUN 1976
C
C
C
C     WCOVAR           VERSION 300     31-OCT-1975
```

```
C
C     THIS ROUTINE DISPLAYS THE CORRELATION MATRIX AFTER
C     CALCULATING IT FROM THE COVARIANCE MATRIX PP.
C
C     IT IS DESIGNED FOR A SYMMETRIC COVARIANCE MATRIX
C     STORED IN THE LOWER LEFT TRIANGLE OF PP WITH
C     VARIANCES STORED IN DIA.  IMAP ASSOCIATES THE
C     MATRIX WITH APPROPRIATE PARAMETER INDICES.
C
      SUBROUTINE WCOVAR(PP,DIA,IMAP,ND)
      DIMENSION PP(11,11),DIA(1),IMAP(1)
      COMMON /IO/LRD,LWD,LSD
C     WRITE(LWD,10)
C  10 FORMAT(' TY 1 FOR CORRELATION MATRIX'/)
C     READ(LRD,15) ICHO
C  15 FORMAT(I)
C     IF (ICHO.NE.1) RETURN
      DO 25 I=1,ND
   25 DIA(I)=SQRT(DIA(I))
C
C     USE IFS AND GOTOS TO HANDLE SPECIAL CASES
C     INSTEAD OF USING DO LOOPS
C
      I=1
   30 IF (I.GT. ND) GOTO 50
      J=1
   35 IF (J.GT.I-1) GOTO 40
      PP(I,J)=PP(I,J)/(DIA(I)*DIA(J))
      J=J+1
      GOTO 35
   40 I=I+1
      GOTO 30
   50 XDIA=1.
      WRITE(LWD,60) (IMAP(I),I=1,ND)
   60 FORMAT('0I ',11(3X,I3,4X)/)
      WRITE(LWD,70) IMAP(1),XDIA
   70 FORMAT(1X,I2,11E10.3/)
      IF (ND.LE.1) GOTO 100
      DO 90 I=2,ND
      I1=I-1
   90 WRITE(LWD,70) (IMAP(I),(PP(I,J),J=1,I1),XDIA)
  100 WRITE(LWD,110)
  110 FORMAT(1X,//)
      RETURN
      END
C     MS2SUB.F4       VERISON 302     19 JAN-1976
C
C     CONTAINS MAP, PARMIO
C     MAP FIXED TO FIX PARAMETERS PROPERLY
C
C     MAP
C
```

```
C     THIS SUBROUTINE TAKES THE ARRAY IA(NP) OF PARAMETERS TO BE FIXED
C     AND SEARCES IT FOR EACH OF THE NP PARAMETERS TO  FORM THE ARRAY
C     IMAP(ND) OF PARAMETERS TO BE VARIED.
C
C
      SUBROUTINE MAP(IA,NP,IMAP,ND)
      DIMENSION IA(1),IMAP(1)
      LOGICAL TEST
      ND=NP
      DO 50 I=1,NP
      IF (IA(I).EQ.0) GOTO  60
   50 ND=ND-1
   60 IM=0
      NF=NP-ND
      DO 100 I=1,NP
      TEST=.TRUE.
      DO  80 J=1,NF
      IF (IA(J).NE.I) GOTO  80
      TEST=.FALSE.
      GOTO  90
   80 CONTINUE
   90 IF (.NOT.TEST) GOTO 100
      IM=IM+1
      IMAP(IM)=I
  100 CONTINUE
      RETURN
      END
C
C
C
C     PARMIO          VERSION 304 28 MAY 1976
C
C     MODIFIED FOR BATCH NNNGER
C
C
C     SUPPORTS SAME PARAMETERS
C
C     PARMIO IS THE BASIC PARAMETER INPUT ROUTINE USED BY
C     OPTION 2,GUESS, 4,FINGR AND 7,ERROR.  IT DETERMINES
C     WHICH OF THE FUNCTIONS WILL BE USED (IWHAT), THE PARAMETER
C     VALUES, THE NUMBER OF PARAMETERS AND THE NUMBER OF COMPONENTS
C     BY CALLING COUNT FOR IWHAT=1,4,5.
C
C
      SUBROUTINE PARMIO(P,NP,NC,NPMAX,IWHAT,NINT,RMSQU,DLMXQU,NIT,*)
      DIMENSION P(1)
      COMMON /IO/LRD,LWD,LSD
C
C     CHOOSE FUNCTION TO BE FIT OR PLOTTED.  IF OSAME CHOSEN,
C     JUMP OUT WITH SAME PARAMETERS AND SAME FIXED PARAMETERS.
C
    5 WRITE(LWD,10)
```

```
   10 FORMAT(' FUNCTION (-1HELP), ITERATIONS, RMS QUIT, DELMX QUIT, (NIT
     1)'/)
      READ(LRD,15) ICHO,NINT,RMSQU,DLMXQU,NIT
*-------------------------
C  15 FORMAT(I2,I3,2E10.2,I5)
   15 FORMAT(2I,2F,I)
      IF (ICHO.LT.0) GOTO 17
      GOTO 19
   17 WRITE(LWD,18)
   18 FORMAT(' 0SAME,1FINGER,2WHATOR,3NUFORM,4EXCESS,5WHTCMP'/)
      GOTO 5
   19 IF (NIT.LE.0) NIT=10
      IF (ICHO.EQ.0) RETURN 1
      IWHAT=ICHO
      IF ((IWHAT.LT.1).OR.(IWHAT.GT.5)) GOTO 5
      GOTO (20,25,30,20,20),IWHAT
   20 WRITE(LWD,21)
   21 FORMAT(' FINAL,FRACT(1),K(1),FRACT(2),K(2),FRACT(3),K(3)'/)
      READ(LRD,22) (P(I),I=1,NPMAX)
*--------------------
C  22 FORMAT(8E10.3)
   22 FORMAT(11F)
      CALL COUNT(P,NC,NPMAX,2)
      NP=2*NC+1
      GOTO 35
   25 WRITE(LWD,26)
   26 FORMAT(' FINAL,FRACT,K,EXP'/)
      READ(LRD,22) (P(I),I=1,4)
      NP=4
      NC=1
      GOTO 35
   30 WRITE(LWD,31)
   31 FORMAT(' FINAL,FRACT,K:TRACER-1,F,K:TRACER-2,DRIVER K,0.44'/)
      READ(LRD,22) (P(I),I=1,7)
      NP=7
      NC=1
      GOTO 35
   35 CONTINUE
      RETURN
      END
C
C
C
C     MISSUB.F4       VERSION 301     27 OCT-1976
C     CORRECTED FOR 370 FORTRAN
C
C     CONTAINS SORTI,COUNT
C
C     SORT: A BUBBLE SORT SORTING A AND Y USING A.
C
C
      SUBROUTINE SORTI(A,N,Y)
```

```
      DIMENSION A(1),Y(1)
      LOGICAL KEY
      IF (N.LE.1) GOTO 30
      N1=N-1
      DO 20 J=1,N1
      L=N-J
      KEY=.TRUE.
      DO 10 K=1,L
      IF (A(K).LE.A(K+1)) GOTO 10
      ATEM=A(K)
      A(K)=A(K+1)
      A(K+1)=ATEM
      YTEM=Y(K)
      Y(K)=Y(K+1)
      Y(K+1)=YTEM
      KEY=.FALSE.
   10 CONTINUE
      IF (KEY) GOTO 30
   20 CONTINUE
   30 RETURN
      END
C
C
C
C     COUNT: COUNTS THE NUMBER OF PARAMETERS.  NOTE EITHER
C     A 0. OR NEGATIVE PARAMETER IS CONSIDERED THE END OF THE
C     PARAMETERS.
C
C
      SUBROUTINE COUNT(P,NC,NM,NPC)
      DIMENSION P(1)
      NC=0
      DO 10 I=2,NM,NPC
      IF (P(I)) 30,30,20
   20 NC=NC+1
   10 CONTINUE
   30 RETURN
      END
C     PLMLT.F4        VERSION 303     29 OCT-1975
C
C     THIS SUBROUTINE PLOTS A LINE ON A TTY OR LPT PUTTING SYMBOLS OF
C     LOCATIONS SPECIFIED IN THE ARRAY R AND USING P,Q AS SCALE
C     FACTORS. IN GENERAL P=100,Q=0.
C
C     Y IS THE ABSCISSA VALUE PLOTTED OUT
C     N IS THE NUMBER OF POINTS TO BE PLO\TTED.
C     NW IS THE WIDTH OF THE PLOT
C     IR IS AN ARRAY OF SYMBOLS PLOTTED. THE NUMBER IN R(I) IS
C       PLOTTED WITH SYMBOL IR(I)
C     LUMP(I) IS THE ACTUAL LINE OF SYMBOLS PLOTTED
C
      SUBROUTINE PLMLT(Y,R,N,LF,P,Q)
```

```
C     PART OF FINGER.CMD.  SAVE.
      INTEGER B,C
      DIMENSION LUMP(51),IR(18),R(1)
      COMMON /IO/LRD,LWD,LSD
      DATA NW/51/
      RNDZ(X)=IFIX(X+SIGN(0.5,X))
      DATA IR/1H.,1H-,1H1,1H2,1H3,1H4,1H5,1H0,1HA,1HB,1HC,1HD,1HE,1HF,
     11HG,1HH,1HI,1HI/
      DATA B/1H /
      DATA C/1HI/
      GO TO (100,100,200),LF
C
C     WRITE THE Y AXIS
C
  100 WRITE(LWD,101)
  101 FORMAT('  DATA  FIT   ECOT    0---.1---.2---.3---.4---.5---.6---.7
     1---.8---.9----1')
      RETURN
C
C     BLANK LINE
C
  200 DO 210 I=1,NW
  210 LUMP(I)=B
      JEND=1
      IF (N.LT.1) GOTO 400
      DO 220 IPL=1,N
C
C     KP IS THE POSITION OF THE SYMBOL TO BE PLOTTED ON THE LINE
C
      KP=IFIX((R(IPL)*P+Q)/2+0.5)+1
      IF ((KP.GT.NW).OR.(KP.LE.0)) GOTO 220
      IF (KP.GT.JEND) JEND=KP
C
C     LUMP BEOMES AN ARRAY OF SYMBOLS TOBE PLOTTED AT POSITION KP
C     BLANK OTHERWISE
C
      LUMP(KP)=IR(IPL)
  220 CONTINUE
      LS=RNDZ(Q)
C
C     PUT THE BASELINE ON THE PLOT LAST SO IT IS NOT OVERWRITTEN
C
      LUMP(LS+1)=C
  400 WRITE(LWD,401)R(N),R(2),Y,(LUMP(J),J=1,JEND)
  401 FORMAT (1H ,F5.3,F6.3,E9.2,1X,51A1)
      RETURN
      END
C
C
C
C     DIFSUB.F4       VERSION 302     19 FEB 1976
C
```

```
C     THIS VERSION HAS BEEN EXTENSIVELY TESTED FOR FINGER, EXCESS,
C     WHATOR AND WHTCMP AND SHOULD BE CORRECT
C     FOR THOSE FUNCTIONS.
C
C     TESTED FOR NUFORM AND ERRORS CORRECTED
C
C
C     FOR EACH OF THESE SUBROTINES, IF IFUNC = 2 ONLY THE FUNCTION
C     VALUE IS CALCULATED
C
C     THE EMPIRICAL DERIVATIVES USED IN FRMDIF CAN BE USED WITH
C     ANY FUNCTION WITH NO CHANGE IN THIS DIFFERENTIATING
C     SUBROUTINE
C
C
C     FINDIF:  FINGER PARTIAL DERIVATIVES, USES ANALYTIC
C       DERIVATIVES
C
C
      SUBROUTINE FINDIF(X,Y,P,PART,NP,NC,SUMC,IFUNC)
      DIMENSION P(1),PART(1)
      SUMC=P(1)
      GOTO (5,50),IFUNC
    5 PART(1)=1.
      DO 10 J=1,NC
      JF=2*J
      JK=JF+1
      DF=1.0/(1.0+P(JK)*X)
      DK=P(JF)*DF
      SUMC=SUMC+DK
      PART(JF)=DF
   10 PART(JK)=-DF*DK*X
      RETURN
   50 DO 60 J=1,NC
      JF=2*J
      JK=JF+1
   60 SUMC=SUMC+P(JF)/(1.0+P(JK)*X)
      RETURN
      END
C
C
C
C     WHTDIF:  ANALYTIC PARTIAL DERIVATIVES FOR WHTCMP
C
      SUBROUTINE WHTDIF(X,Y,P,PART,NP,NC,SUMC,IFUNC)
      DIMENSION P(1),PART(1)
      SUMC=P(1)
      GOTO (5,50),IFUNC
    5 PART(1)=1.
      DO 10 J=1,NC
      JF=2*J
      JK=JF+1
```

```
      DF=(1.+P(JK)*X)**(-0.44)
      DK=P(JF)*DF
      SUMC=SUMC+DK
      PART(JF)=DF
   10 PART(JK)=-0.44*X*P(JF)/(1.+X*P(JK))**(1.44)
      RETURN
   50 DO 60 J=1,NC
      JF=2*J
      JK=JF+1
   60 SUMC=SUMC+P(JF)*(1.0+P(JK)*X)**(-0.44)
      RETURN
      END
C
C
C
C     EXCDIF:  ANALYTIC DERIVATIVES FOR EXCESS
C
      SUBROUTINE EXCDIF(X,Y,P,PART,NP,NC,SUMC,IFUNC)
      DIMENSION P(1),PART(1)
      SUMC=P(1)
      GOTO (5,50),IFUNC
    5 PART(1)=1.
      DO 10 J=1,NC
      JF=2*J
      JK=JF+1
      DF=EXP(-P(JK)*X)
      DK=P(JF)*DF
      SUMC=SUMC+DK
      PART(JF)=DF
   10 PART(JK)=-X*DK
      RETURN
   50 DO 60 J=1,NC
      JF=2*J
      JK=JF+1
   60 SUMC=SUMC+P(JF)*EXP(-P(JK)*X)
      RETURN
      END
C
C
C
C     WHADIF:  ANALYTIC DERIVATIVES FOR WHATOR
C
      SUBROUTINE WHADIF(X,Y,P,PART,NP,NC,SUMC,IFUNC)
      DIMENSION P(1),PART(1)
      SUMC=P(1)
      GOTO (5,50),IFUNC
    5 PART(1)=1.
      DO 10 J=1,NC
      JK=3*J
      JF=JK-1
      JE=JK+1
      DF=1/(1.+P(JK)*X)**P(JE)
```

```
      DK=P(JF)*DF
      PART(JF)=DF
      PART(JK)=-P(JE)*P(JF)*X/(1.+P(JK)*X)**(1.+P(JE))
      PART(JE)=-DK*ALOG(P(JF)*(1.+X*P(JK)))
   10 SUMC=SUMC+DK
      RETURN
   50 DO 60 J=1,NC
      JK=3*J
      JF=JK-1
      JE=JK+1
   60 SUMC=SUMC+P(JF)/(1.+P(JK)*X)**P(JE)
      RETURN
      END
C
C
C
C     FRMDIF:  EMPIRICAL DERIVATIVES FOR ANY FUNCTION FORM
C                OF 7 PARAMTERS
C
      SUBROUTINE FRMDIF(X,Y,P,PART,NP,NC,SUMC,IFUNC)
      DIMENSION P(1),PART(1)
      CALL FORM(P,X,SUMC)
      IF (IFUNC.EQ.2) RETURN
      DEL=.01
      DO 10 J=1,NP
      TEMP=P(J)
      IF (TEMP.NE.0.0) GOTO 15
      PART(J)=0.
      GOTO 10
   15 P(J)=TEMP*(1.+DEL)
      CALL FORM(P,X,YVAL1)
      PART(J)=(YVAL1-SUMC)/(DEL*TEMP)
      P(J)=TEMP
   10 CONTINUE
      RETURN
      END
C     NFORM.F4        VERSION 1    18-FEB-1976
C
C
C
C
C     FORM: CALCULATES THE FUNCTION DECRIBED IN NUFORM.
C
C     F(X,P)=FRACTION SINGLE STRANDED TRACER  WITH RATES
C     P(3),P(5) AND FRACTIONS P(2),P(4) DRIVEN BY
C     DRIVER WITH RATE P(6) WITH SINGLE STRAND DRIVER
C     GIVEN BY P(7)=0.44 USUALLY.
C
      SUBROUTINE FORM(P,COT,VAL)
      REAL P(1)
      V=1+P(6)*COT
      PWR=1-P(7)
```

```
VAL1=P(1)
VAL2=P(3)/P(6)*(1-V**PWR)/PWR
VAL2=P(2)*EXP(VAL2)
VAL3=P(5)/P(6)*(1.-V**PWR)/PWR
VAL3=P(4)*EXP(VAL3)
VAL=VAL1+VAL2+VAL3
RETURN
END
```