# Deoxyribonucleic Acid Sequence Organization in the Genome of the Dinoflagellate *Crypthecodinium cohnii*[†]

Alan G. Hinnebusch, Lynn C. Klotz,* Ellen Immergut, and Alfred R. Loeblich III[‡]

ABSTRACT: Details of the general DNA sequence organization in the dinoflagellate *Crypthecodinium cohnii* have been obtained by using hydroxylapatite binding experiments, $S_1$ nuclease digestion, and electron microscopy of reassociated DNA. It has been found that roughly half of the genome is made up of unique sequences interspersed with repeated sequence elements with a period of ~600 nucleotides. This class represents roughly 95% of the total number of interspersed unique elements in the genome. The remaining 5% are uninterrupted by repeated sequences for at least 4000 nucleotide pairs. The interspersed repeated elements are narrowly distributed in length with 80% under 300 nucleotide pairs in length. About half of the repeated DNA (20–30% of the genome) is not interspersed among unique sequences. The close spacing of the short repeats interspersed throughout much of the genome is consistent with the occurrence of the huge network structures observed in the electron microscope for low $C_0t$ reassociation of moderately long fragments. An unusual class of heteroduplexes was detected in the electron microscope which is believed to derive from the reassociation of repeated sequences from different families which are frequently found adjacent to one another in different locations in the genome. The occurrence of this novel arrangement of repeated sequences may reflect the unusual organization of the dinoflagellate nucleus. However, in most respects the sequence arrangement in this unicellular alga is very typical of higher plants and animals.

There are great similarities in the organization of the bulk of the DNA sequences in the genomes of many multicellular eucaryotes, both plants and animals. In these organisms, there exists a fine interspersion of repetitive and fairly nonrepetitive or unique DNA sequences throughout the greater part of the genome. Typically, the nonrepetitive sequence elements are interrupted every 1000–2000 nucleotide pairs by a repetitive sequence of a few hundred nucleotide pairs (Davidson et al., 1975). In several higher plants, the unique sequence elements in this so-called short-period interspersion pattern are even shorter, averaging around 500 nucleotide pairs (Smith & Flavell, 1977; Murray et al., 1978). A large difference in the lengths of the interspersed unique and repeated sequence elements is found in a few insects (Manning et al., 1975; Crain et al., 1976a,b; Wells et al., 1976) and a nematode worm (Beauchamp et al., 1979) where the interspersed elements are at least an order of magnitude larger in size. The organization of the genomes of the few avian species that have been examined appears to be intermediate between these two cases (Arthur & Straus, 1978; Epplen et al., 1978; Eden & Hendrick, 1978).

Both extremes of sequence arrangement occur among the eucaryotic protists. The cellular slime mold *Dictyostelium discoideum* possesses a short-period interspersion pattern involving most of its genome (Firtel & Kindle, 1975), while the water mold achlya, a primitive fungus, displays long-period interspersion (Hudspeth et al., 1977). Two representatives of the higher fungi, *Aspergillus nidulans* and *Saccharomyces cerevisiae*, appear to contain little or no repeated DNA whatsoever beyond the rRNA coding gene families (Timberlake, 1978; Lauer et al., 1977). The same is true of the unicellular green algae *Chlamydomonas reinhardtii* (Howell & Walker, 1976) and *Chlorella pyrenoidosa* (Bayen & Dalmon, 1975). Clearly, the evolutionary relationship between different forms of DNA sequence arrangement is not obvious.

The dinoflagellates are a group of diverse eucaryotic protists, generally classified with the eucaryotic algae, possessing a number of peculiar traits of nuclear organization suggestive of the procaryotic state. Among these are (1) the absence of detectable histones and the procaryotic dimensions of the chromatin fiber (55–60 Å), (2) the ultrastructure of the chromatin fibrils, which occur in arched whorls similar to those visible in bacterial nucleoids, (3) permanent condensation of the chromosomes throughout the cell cycle, and (4) the absence of a mitotic spindle and the role of chromosomal membrane attachments in the segregation of daughter chromosomes (Loeblich, 1976; Hamkalo & Rattner, 1977; Livolant & Bouligand, 1978). These procaryotic affinities of the dinoflagellates could indicate their closeness to the ancestral state of all present day eucaryotes, at least regarding nuclear organization (Loeblich, 1976).

In view of this possibility, we have examined the details of DNA sequence arrangement in the free-living, nonphotosynthetic dinoflagellate *Crypthecodinium cohnii.* Earlier work (Allen et al., 1975) indicated the occurrence of a large fraction of the genome of *C. cohnii* as repeated DNA sequence and suggested a fine interspersion of these repetitive sequences among the nonrepeated DNA. We wished to confirm these preliminary indications and to extend the analysis of sequence arrangement in *C. cohnii* by using the methods developed by Britten and Davidson (Davidson et al., 1973) which have been applied by many workers to the various groups of organisms described above.

## Materials and Methods

*DNA Isolation. C. cohnii* DNA was isolated by using a modification of the method of Blin & Stafford (1976). Late

[†] From the Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138. *Received August 20, 1979; revised manuscript received January 8, 1980.* This work was supported by grants from the National Institutes of Health to A.R.L. and L.C.K. and from the National Science Foundation to L.C.K. A.G.H. was supported by a predoctoral graduate fellowship from the National Science Foundation.

* Address correspondence to this author at the Department of Biochemical Sciences, Princeton University, Princeton, NJ 08544.

[‡] Present address: University of Houston, Marine Science Program, Galveston, TX 77550.

0006-2960/80/0419-1744$01.00/0

log-phase cultures grown at 27 °C in sterile MLH (Tuttle & Loeblich, 1975) were used in all cases. (Cells from stationary phase cultures could not be lysed by using the procedures outlined below.) Cells were harvested at 2500$g$ for 5 min at 0 °C. The pellets were resuspended and consolidated at 4 °C in 0.5 M sodium–EDTA (pH 8), 0.5% sarkosyl, and 100 $\mu$g/mL proteinase K. After the cells were pelleted a second time, the supernatant was placed in a 50 °C bath. The cell pellet was frozen in liquid $N_2$ and ground with a mortar and pestle to a fine powder. The powder was returned in small portions to the reserved supernatant and resuspended by swirling. The whole mixture was then agitated for 2 to 3 h on a rotary shaker at 50 °C until the ground cells were generally dispersed. A rubber policeman was used to aid this procedure. The lysate was then made 7 mM in $\beta$-mercaptoethanol and 10 $\mu$g/mL in ethidium bromide, and KI was added to make the refractive index 1.439. This solution was centrifuged at 10000$g$ for 30 min at 5 °C, after which the liquid layer was removed from beneath a precipitate that had formed on top. This solution was then centrifuged to equilibrium at 40000 rpm at 20 °C. The DNA bands were visualized under long-wave UV and collected from the sides of the tubes by using small gauge syringe needles. Ethidium bromide was extracted 3 times with isoamyl alcohol, and the DNA was dialyzed into 0.3 M NaCl and 0.03 M sodium citrate. After dialysis, heat-treated (85 °C, 10 min) RNase was added to 100 $\mu$g/mL and incubated at 37 °C for 2 h, followed by a 2-h proteinase K digestion at 150 $\mu$g/mL. KI was added to make the index of refraction 1.428. Mercaptoethanol and ethidium bromide were included, and the solution was centrifuged to equilibrium for a second time, as before. The DNA was collected and the ethidium bromide extracted, as above, and then the solution was dialyzed into 0.01 M Tris-HCl (pH 7.4) and 1 mM EDTA for storage. The 260 nm/280 nm and 260 nm/230 nm absorbance ratios of this DNA were always in excess of 1.8 and 2.1, respectively. Values for the $T_m$ and buoyant density were determined and agree with published values (Rae, 1973; Allen et al., 1975).

Unlabeled *C. cohnii* DNA was isolated from a wild type strain already described (Allen et al., 1975). Tritium-labeled DNA was prepared in vivo by using an adenine auxotroph of *C. cohnii* isolated by Tuttle & Loeblich (1977). Growth of the mutant was carried out in MLH supplemented with 10 $\mu$g/mL adenine. The last two doublings before harvesting occurred in the presence of 5 $\mu$Ci/mL [$^3$H]adenine (15–30 Ci/mmol).

Isolation of *Escherichia coli* and T7 DNAs has already been described (Hinnebusch et al., 1978). Calf thymus DNA and yeast tRNA were purchased from Sigma. The pBR322 and $\phi$X174 DNAs were the gifts of Drs. H. Lehrach and K. Koths, respectively.

*Preparation of Sheared DNA Samples.* DNA was sheared to various average lengths with a Virtis 60 homogenizer (Britten et al., 1974). After being sheared, the DNA was fractionated on 13-mL 5–20% alkaline sucrose gradients made in 0.9 M NaCl, 0.1 M NaOH, and 1 mM EDTA. For the preparation of a single size class from a sheared sample, the center 50% of the distribution was pooled. In the isolation of radioactive tracers of various lengths, the peak fractions were pooled into five or six different samples. Average single-strand lengths of the prepared samples were determined as already described (Hinnebusch et al., 1978) by using alkaline agarose gel electrophoresis and T7/*Hpa*I restriction fragments as size markers.

*DNA Reassociation Analysis.* Samples used in reassociation reactions were dialyzed into solutions that had been treated with a chelating ion-exchange resin (Chelex 100, Bio-Rad) to remove divalent metal ions. DNA concentrations were determined by measuring the optical density at 260 nm. Denaturation was accomplished in one of two ways. For the short fragment self-driven reactions in Figure 1 (○), the DNA was dialyzed into 0.12–1.0 M PB[1] and denatured by heating for 10 min at 100–110 °C [the denaturation temperature was adjusted for the dependence of the DNA melting temperature on salt concentration (Schildkraut & Lifson, 1965)]. The sample was returned to a temperature 25 °C below the melting temperature for that salt concentration to initiate renaturation. For all other reactions, heat denaturation was avoided because of its degradative effect on large DNA fragments (unpublished experiments) and pH denaturation was employed instead. In these reactions, the DNA was dialyzed into unbuffered NaCl of the appropriate concentration and denaturation was accomplished by adding 1 part 1 N NaOH to 19 parts DNA and incubating at room temperature for 10 min. For HA-monitored kinetics, 1 part 2 M $NaH_2PO_4$ was added to initiate renaturation. For $S_1$ nuclease assayed reactions, 0.5 part 1 M Tris-HCl (pH 7.4) and 1 part 1 M HCl were used for neutralization since PB inhibits the activity of $S_1$ nuclease. The equivalent $C_0t$ values were calculated for each reaction by using the known dependence of DNA renaturation rate on cation concentration (Britten et al., 1974).

HA assay of reassociation was carried out exactly as previously described (Hinnebusch et al., 1978) by using a temperature elution (100 °C, 0.12 M PB) of the double-stranded fractions. In low $C_0t$ reactions involving small amounts of DNA, carrier DNA (20 $\mu$g of $C_0t$ = 7 mol $L^{-1}$ s renatured, sheared calf thymus DNA) was added to the samples before chromatography on HA. $S_1$ nuclease digested samples were assayed on Sephadex G-100 (Smith et al., 1975). Bed volumes of at least 20 times the sample size were used, and the volume of each fraction was 0.1 the bed volume. Undigested DNA elutes at the void volume ($V_0$), while the digestion products of denatured DNA elute at $\sim 2.5 V_0$. The radioactivity in each fraction was determined by liquid scintillation and the amount eluting between $K_{av}$ = 0 and $K_{av}$ = 0.5 was summed to calculate the fraction of a sample undigested. The ratio of this value to the total radioactivity recovered was taken to be the fraction undigested.

The same general procedure used in assaying reassociation reactions with HA was employed in HA thermal chromatography, with the collection of three 0.12 M PB washes at each temperature, each equal to 2 bed volumes in size. The column temperature during elution was calculated from the circulating bath temperature by using an empirically determined relationship between the two.

Large networks in reassociated samples were detected by their ability to bind to HA in 0.40 M PB at 60 °C. Elution was carried out at 100 °C in 0.40 M PB. The [PB] in each fraction was adjusted to 0.20 M before adding 6.5 volumes of scintillation fluid in order to accurately measure $^3$H radioactivity in all of the fractions.

*Characterization of Reassociated Calf DNA.* Calf DNA was labeled in vitro by nick translation at 15 °C (Maniatis et al., 1975) and then fractionated on an alkaline sucrose gradient as described above. Sheared unlabeled calf DNA was

[1] Abbreviations used: PB, sodium phosphate buffer, pH 6.9; HA, hydroxylapatite; $C_0t$, the product of DNA phosphate concentration in moles per liter and time in seconds; $Cl_3AcOH$, trichloroacetic acid; $\bar{L}_w$ and $\bar{L}_n$ are weight- and number-average fragment lengths, respectively.

sedimented in a parallel gradient. Equivalent fractions were pooled in the two gradients to give symmetrical peaks of relatively high molecular weight. The labeled and unlabeled DNAs were found to have $\bar{L}_w$ values of 1340 and 1400 nucleotides, respectively. The zero-time binding fraction (37% at $C_0t < 1 \times 10^{-4}$ mol $L^{-1}$ s) was assumed to be artifactually large due to the labeling procedure and was removed on HA. The fraction binding to HA after reassociation to $C_0t = 5$ mol $L^{-1}$ s after removal of this fraction was 57%, which compares favorably with published data (Britten & Smith, 1970).

*$S_1$ Nuclease Digestions.* The $S_1$ nuclease used in these experiments was purchased from Sigma. Reactions were carried out in 0.025 M $KCH_3COO$ (pH 4.5), 1 mM $ZnSO_4$, and 0.5 M NaCl at 45 °C. The DNA concentration in the digestions was variable and will be indicated as the results are discussed. The extent of digestion was determined either by gel filtration on Sephadex G-100 (see above) or by $Cl_3AcOH$ precipitation. The enzyme was assayed by using heat-denatured sheared $^3$H-labeled *E. coli* DNA at 10.4 μg/mL in 0.5-mL reactions under the standard conditions described above. Conversion of this DNA to acid solubility was found to be linear for the first 50% of the reaction from which the activity was calculated. One unit is defined as the amount of $S_1$ nuclease capable of converting 1 μg of this DNA to acid solubility per min. The $S_1$ nuclease preparation had a concentration of 2.1 units/μL.

*Sizing of $S_1$ Nuclease Resistant Duplexes.* $S_1$ nuclease resistant DNA was isolated by ethanol precipitation. Control experiments were done to ensure that the digestion products were not precipitable under the salt and DNA concentrations employed. The results (not presented) indicated that less than 0.1% of a Sephadex G-100 included fraction ($K_{av} > 0.5$) was found in a washed precipitate formed by tRNA added to the digested sample as carrier. Precipitated samples were resuspended in half-strength electrophoresis buffer, which at full strength was 40 mM Tris-acetate (pH 7.8), 5 mM $NaCH_3COO$, and 1 mM EDTA for agarose gels and 50 mM Tris-borate (pH 8.3) and 1 mM EDTA for polyacrylamide gels. After electrophoresis, agarose gels were stained and sliced and the slices were assayed for radioactivity as already described (Hinnebusch et al., 1978). For polyacrylamide gels, the gels were cast with 10% glycerol to facilitate slicing of the frozen gels. The detection of radioactivity in the slices required the use of 3% Protosol and 0.4% Omnifluor (New England Nuclear) in toluene. The slices were incubated in this mixture at room temperature, with occasional agitation, until they became transparent and were then counted.

*Electron Microscopy of Reassociation Products.* The formamide-basic protein technique of Davis et al. (1971) was used for all samples. The hyperphase was 40% (v/v) formamide, 0.1 M Tris-HCl (pH 8), and 0.01 M EDTA. Melting curve analysis on unsheared *E. coli* DNA dialyzed into the hyperphase solution revealed a reduction of the $T_m$ for this DNA of 28 °C from its value in 0.12 M PB (data not shown). Thus, the equivalent of the reassociation temperature in the hyperphase buffer is 32 °C (60 °C − 28 °C). This indicates that at room temperature the reassociation products will be stable under the spreading conditions used. The DNA was picked up on parlodian-coated grids, stained with uranyl acetate, and rotary shadowed with 60% platinum-20% palladium. The grids were examined in a Phillips 300 electron microscope, and pictures were taken at random from locations on the grids where strand density and contrast were acceptable. All pictures were taken at a magnification of 16000× and enlarged in printing 3.6× to give a final magnification of 57600×. Contour lengths were measured with a Wang 2200 minicomputer equipped with a digitizer.

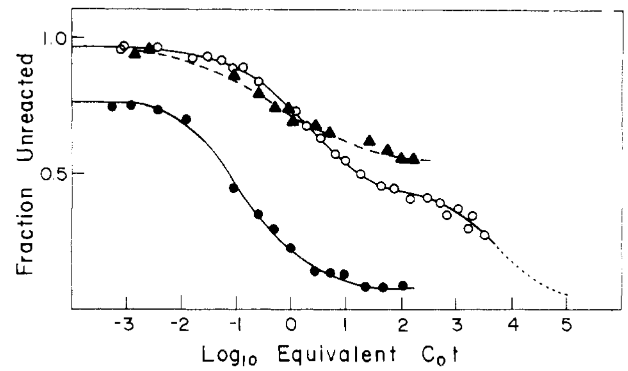

FIGURE 1: Reassociation kinetics of $^3$H-labeled *C. cohnii* DNA at two different fragment lengths. (○) HA-assayed reassociation of fragments 260 nucleotides long. The solid line through the data up to log $C_0t = 2.65$ merely connects the points. The line drawn through the higher $C_0t$ points (including the dotted portion) is the calculated second-order curve for the unique fraction based on the haploid DNA content (see text) and assuming that 37% of the fragments are wholly unique and that 95% of the fragments are reactable (based on the average reactability of *E. coli* DNA fragments of the same length; data not presented). (●) HA-assayed reassociation of 3500-nucleotide fragments. The curve shown is hand drawn to merely connect the points. (▲) Reassociation of 3500-nucleotide fragments assayed for resistance to $S_1$ nuclease digestion. Digestions were carried out at [DNA] = 12 μg/mL and [$S_1$ nuclease] = 3.6 units/mL.

## Results

*Self-Driven Reassociation Kinetics at Two Fragment Lengths.* Allen et al. (1975) have described the HA reassociation kinetics of short fragments of *C. cohnii* DNA between the $C_0t$ values of 0.05 and 3000 mol $L^{-1}$ s. Our data shown in Figure 1 (○) from an HA-assayed reassociation of [$^3$H]-DNA fragments of $\bar{L}_w = 260$ nucleotides confirm their results and extend the analysis to a $C_0t$ value of $10^{-3}$ mol $L^{-1}$ s. The repeated sequences renature over 4 to 5 orders of magnitude of $C_0t$, indicating a broad spectrum of reiteration frequencies, ranging from fewer than 100 copies to as many as $10^5$ copies per haploid genome. The highest $C_0t$ transition occurs at the rate expected for single-copy sequences based on a haploid DNA content of 3.8 pg (Allen et al., 1975). At the lowest $C_0t$ values attainable, $<10^{-4}$ mol $L^{-1}$ s, ~4% of the denatured fragments were bound to HA (data points not shown). This "zero-time" binding has been observed for all eucaryotic DNAs examined and has been attributed to the occurrence of inverted-repeated sequences in the eucaryotic genome (Wilson & Thomas, 1974; Perlman et al., 1976).

Also plotted in Figure 1 (●) are the results of a second HA-assayed reassociation using much larger fragments with $\bar{L}_w = 3500$ nucleotides. It can be seen that the increase in fragment length has been accompanied by an increase in the extent of the reaction attributable to repeated sequences. This increase is due in part to the greater zero-time reaction which is 25% at this larger fragment length. However, if we subtract out its contribution by using the correction suggested by Davidson et al. (1973) (see Figure 2 legend), we find that the binding at low $C_0t$ (<100 mol $L^{-1}$ s) has increased from 58 to 91%. This is an indication of an interspersion of repeated and nonrepeated DNA sequences in the *C. cohnii* genome, with a period of less than 3500 nucleotide pairs (Davidson et al., 1973). If this is so, the HA-bound reassociation products formed on large fragments should be extensively single stranded near the end of the repeated sequence reassociation. That this is true is suggested by the results of the $S_1$ nuclease assayed reassociation of the 3500 nucleotide long fragments, also shown in Figure 1 (▲). When more than 90% of the fragments are HA bindable at $C_0t$ = 100 mol L$^{-1}$ s, only 45% of the DNA nucleotides are $S_1$ nuclease resistant. $S_1$ nuclease digestion of reassociated *C. cohnii* repeated sequences will be examined in greater detail below.

*Measurement of the Interspersion Period.* The increase in HA binding with increasing fragment length due to repeated sequence reassociation can be used to estimate the average spacing between repeated sequence elements (Davidson et al., 1973). Briefly, radioactive tracers of different lengths are driven to reassociate by short unlabeled fragments, present in excess, to a driver-determined $C_0t$ value at which largely only repeated sequences in the driver will have reassociated. The short driver approach eliminates the problem of estimating the length dependence of the rate of reassociation of repeated sequences whose organization is unknown. However, it does not eliminate the problem of the increased rate of the unique sequence fraction with increasing tracer length in a driver–tracer reaction (Hinnebusch et al., 1978; Chamberlin et al., 1978). For this reason, we chose to remove fragments containing only unique DNA sequences from the driver before using it in the driver–tracer reactions (Manning et al., 1975). This was accomplished by renaturing fragments 270 nucleotides long to $C_0t$ = 60 mol L$^{-1}$ s and isolating the fraction of this DNA which bound to HA (55%, as expected from Figure 1). After this bound fraction was thermally eluted from HA, its average length was essentially unchanged. At $C_0t$ = 60 mol L$^{-1}$ s, only slightly more than 1% of the unique DNA sequences are expected to have reassociated while 92% of the repeated sequences will have reacted. Thus, this driver is almost totally free of nonrepeated sequences but may be slightly underrepresentative of the most slowly reassociating repeat families. However, since the driver–tracer reactions are carried out to apparent completion (see below), this loss is probably insignificant.

Twelve out of the 18 tracers to be used were sized in duplicate. The percent standard errors on these replicate determinations ranged from 1 to 5%. (Before proceeding further, we measured the amount of zero-time binding for each tracer. These values were found to be concentration independent for all of the tracers in the $C_0t$ range of $3 \times 10^{-5}$–$1.25 \times 10^{-4}$ mol L$^{-1}$ s and are plotted in Figure 2A as a function of the number-average tracer length.) The driver–tracer reactions were carried out in all cases to $C_0t$ = 80 mol L$^{-1}$ s, although the reactions appeared to be complete by $C_0t$ = 20 mol L$^{-1}$ s. The fractions of the tracer sequences bound to HA in these reactions are plotted in Figure 2B (○) vs. the number-average tracer lengths. They are corrected for the amount of zero-time binding of the denatured tracers (see Figure 2 legend). Also plotted in Figure 2B on the ordinate (▲) is our estimate of the fraction of the *C. cohnii* genome in repeated sequence, obtained from hyperchromicity and $S_1$ nuclease digestion measurements on reassociated repeated sequences (presented below). This interpretation of the ordinate intercept is based on a theoretical analysis of the HA binding curve for a simplified genome arranged as interspersed unique and repeated sequence elements (Graham et al., 1974). From the figure, it is evident that the binding increases substantially for fragment lengths up to ~1000 nucleotides. From this length out to at least 4000 nucleotides, the increase in HA binding with increasing tracer length is decreased. In addition, the binding curve in this length range is linear. This indicates that the great majority of the interspersed single-copy sequence elements are less than 1000 nucleotide pairs in length and that a minor fraction are greater than 4000 nucleotide pairs. There are very few, if any, of intermediate length. A more complete analysis of these binding data will be presented under Discussion.

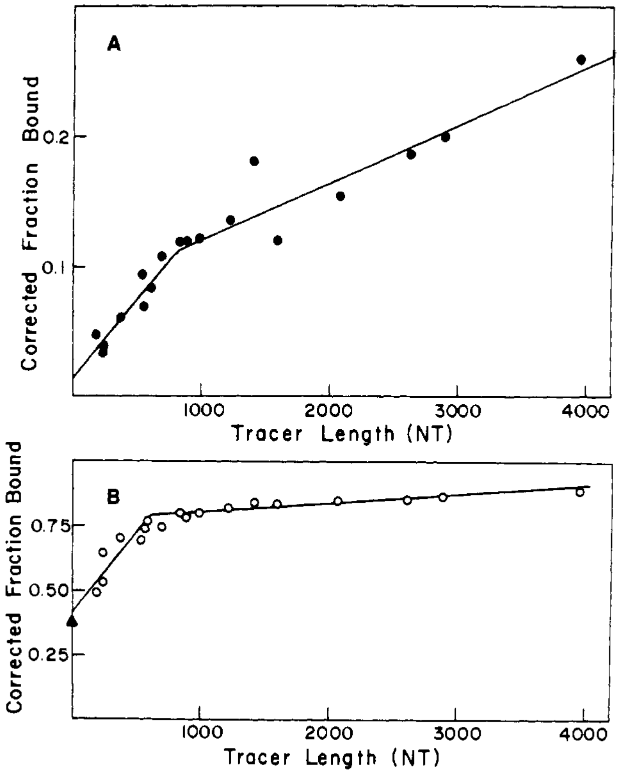

FIGURE 2: Binding of reassociated *C. cohnii* DNA tracers to HA as a function of the number-average tracer length in nucleotides (NT). (A) Zero-time reactions. The curve shown is hand drawn through the data. (B) Driven reactions. Tracers were reassociated in the presence of a 100-fold mass excess of "repeated sequence" driver. For each tracer, the fraction bound to HA in the driven reaction (*B*) was corrected for the zero-time reaction (*Z*) according to $H(L) = (B - Z)/(1 - Z)$ (Davidson et al., 1973). The value of $H(L)$ is plotted for each tracer. The solid line drawn represents a "best-fit" in a least-squares sense to a set of equations derived by Schmid & Deininger (1975).

*Measurement of the Fraction of the Genome in Repeated DNA Sequences.* (*A*) *Hyperchromicity Measurements.* The hyperchromicity of reassociated DNA can be used to estimate the fraction of the DNA which is actually base-paired (Davidson et al., 1973). We have carried out this measurement on reassociated *C. cohnii* DNA by using fragments of $\bar{L}_w$ = 3500 nucleotides reassociated to $C_0t$ = 50 mol L$^{-1}$ s and fragments of $\bar{L}_w$ = 2400 nucleotides reassociated to $C_0t$ = 100 mol L$^{-1}$ s. The reassociation of the repeated sequences is expected to be more than 90% complete in these reactions, while the unique sequence fraction at this fragment length may be as much as 5% reacted. Thus, the fraction of the DNA base-paired in these samples is a good estimate of the fraction of the genome in repeated DNA sequences. The results for the two samples are very similar and are shown in Figure 3A along with a melting curve on 250 nucleotide pair native fragments of *C. cohnii* purified on HA beforehand to remove small or denatured fragments. The melting curves of the reassociated fragments are very broad in comparison to the native fragment curve, suggesting a wide range of thermal stabilities among the reassociation products. Some of this duplex melts at temperatures just above the renaturation temperature (60 °C in 0.12 M PB). The $T_m$ of the reassociated duplex is estimated from these melts at 72.2 ± 0.2 °C, which is reduced by 9.0 °C from the $T_m$ of the short native fragments. This suggests that the reassociated duplex is ~10%

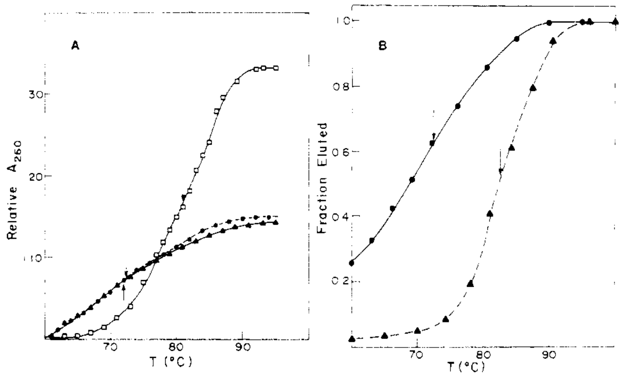

FIGURE 3: (A) Optical melting curves on sheared native and reassociated *C. cohnii* DNA: (●) $C_0t$ = 100 mol $L^{-1}$ s reassociated fragments of $\bar{L}_{\rm w}$ = 2400 nucleotides; (▲) $C_0t$ = 50 mol $L^{-1}$ s reassociated fragments of $\bar{L}_{\rm w}$ = 3500 nucleotides; (□) sheared double-stranded fragments of $\bar{L}_{\rm w}$ = 250 nucleotide pairs. The arrows indicate the $T_{\rm m}$ values for each experiment (see text). All three experiments were carried out in 0.12 M PB. The calculation of the fraction base-paired in the reassociated samples follows Graham et al. (1974): $(H_{\rm R} - 0.04)/[H_{\rm N}(1 - m) - 0.04]$, where $H_{\rm R}$ is the measured hyperchromicity of the reassociated sample, $H_{\rm N}$ is the hyperchromicity of the native sample, $m$ is the fraction of the base pairs mismatched in the reassociated sample as estimated from its $T_{\rm m}$, and 0.04 is the hyperchromicity of single-stranded DNA. For *C. cohnii* DNA, $m$ is given the value of 0.10 (see text). (B) Thermal chromatography on HA of sheared native fragments and reassociated fragments of *C. cohnii* DNA: (●) elution profile of a $C_0t$ = 80 mol $L^{-1}$ s reassociated sample of a tracer of $\bar{L}_{\rm w}$ = 560 nucleotides driven by a 100-fold mass excess of repetitive driver; (▲) thermal elution profile of native fragments sheared to $\bar{L}_{\rm w}$ = 400 nucleotide pairs. The arrows indicate the $T_{\rm m}$ values (see text).

mismatched (Britten et al., 1974).

From the ratio of the relative hyperchromicities of the reassociated and native samples, we have calculated the fraction of the nucleotides that are base-paired in the reassociation products (see legend to Figure 3 for details). The resulting values are 0.39 and 0.42 for the $C_0t$ = 50 and 100 mol $L^{-1}$ s reactions, respectively. The hyperchromicity of the sheared native DNA shown in Figure 3A is ~10% lower than the value observed prior to shearing. Since it is not clear which value is the best to use in making the above calculation, we report our estimates for the fractions base-paired in the $C_0t$ = 50 and 100 mol $L^{-1}$ s samples as 0.33–0.39 and 0.36–0.42, respectively.

These estimates agree with that of Allen et al. (1975), who measured the hyperchromicity of HA-isolated $C_0t$ = 100 mol $L^{-1}$ s reassociated *C. cohnii* duplex formed on shorter (500–600 nucleotides) fragments. However, the $T_{\rm m}$ reported there was 78.3 °C, which is significantly larger (6 °C) than the value reported here. In an attempt to verify our results, we have performed thermal chromatography on HA of $C_0t$ = 80 mol $L^{-1}$ s reassociated *C. cohnii* DNA. In order to avoid the formation of network structures (see below), which give spurious HA thermal elution profiles (Thompson, 1976), we reassociated one of the smaller tracers, described in the last section, having $\bar{L}_{\rm w}$ = 560 nucleotides, with an excess of the short repetitive driver. By use of an HA assay for network formation (Flavell & Smith, 1977; see Materials and Methods), it was shown that large network structures were not formed in this reassociation reaction. The thermal elution profiles for this sample and a native sheared sample of *C. cohnii* DNA with $\bar{L}_{\rm w}$ = 400 nucleotide pairs are shown in Figure 3B. The $T_{\rm m}$ measured for the native fragments (▲) was 82 °C, within 1 °C of the expected $T_{\rm m}$ in an optical melt on fragments of this length (Britten et al., 1974). The $T_{\rm m}$ of the reassociated DNA (●) was 72.3 °C, in good agreement with the values observed in the optical melts on the $C_0t$ = 50 and 100 mol $L^{-1}$ s reassociated samples shown in Figure 3A. We conclude that the $T_{\rm m}$ of the reassociated repeated sequences formed at $T_{\rm m}$ = 25 °C is 10 ± 1 °C below that of native *C. cohnii* fragments at a size comparable to the reassociated repeats (see below).

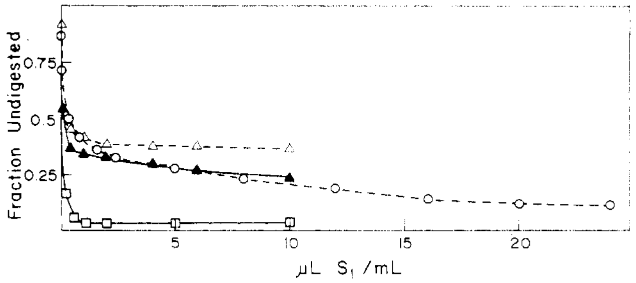

FIGURE 4: $S_1$ nuclease digestion kinetics: (○) $C_0t$ = 100 mol $L^{-1}$ s reassociated *C. cohnii* DNA of $\bar{L}_{\rm w}$ = 3500 nucleotides carried out at 12.4 μg/mL; (▲) $C_0t$ = 5 mol $L^{-1}$ s calf DNA of $\bar{L}_{\rm w}$ = 1340 nucleotides carried out at 10 μg/mL; (△) $C_0t$ = 1.8 mol $L^{-1}$ s reassociated *E. coli* fragments of $\bar{L}_{\rm w}$ = 3000 nucleotides at 16.5 μg/mL; (□) denatured sheared *E. coli* fragments at 10.4 μg/mL. The *C. cohnii* digestion was assayed on Sephadex G-100, and the other digestions were assayed by $Cl_3AcOH$ precipitation. The $S_1$ nuclease concentration was 2.1 units/μL.

(*B*) *$S_1$ Nuclease Digestion Measurements.* An alternative approach to measuring the fraction of the genome in repeated DNA sequences is provided by $S_1$ nuclease digestion of the reassociation products, as already described in Figure 1. The validity of this approach depends upon the specificity of the enzyme; it must be shown that the nuclease digests long single-stranded regions in reassociation products at a much higher rate than it digests short un-base-paired sites in the duplex regions formed in the reassociation of mismatched repeated sequences. We have examined the digestion kinetics of reassociated repeated *C. cohnii* DNA to determine if this requirement is met for this DNA. These data are shown in Figure 4 (○) for fragments of $\bar{L}_{\rm w}$ = 3500 nucleotides reassociated to $C_0t$ = 100 mol $L^{-1}$ s. Although a simple limit digestion comparable to that seen with partially reassociated *E. coli* DNA fragments (△) is not observed, a large change in the rate of digestion occurs at the enzyme concentration at which fully single-stranded DNA becomes completely acid soluble (□). The same is true for $C_0t$ = 5 mol $L^{-1}$ s reassociated calf DNA fragments of $\bar{L}_{\rm w}$ = 1340 nucleotides (▲), although in this case the change in rate is reproducibly more abrupt. These kinetics have been determined in duplicate for calf DNA and repeatedly for *C. cohnii* with the same results, regardless of whether $Cl_3AcOH$ precipitation or exclusion from Sephadex G-100 is used to assay resistance to digestion or whether digestion is carried out in 0.5 M $Na^+$ at 45 °C or 0.2 M $Na^+$ at 15 °C.

We interpret these results to indicate that the first phase of the reaction involves, primarily, the digestion of long single-stranded regions of unreassociated DNA. The second phase involves the random nicking at sites of mismatch, involving only one or a few unpaired bases. $S_1$ nuclease is known to catalyze such a reaction at high enzyme concentrations (Shenk et al., 1975; Dodgson & Wells, 1977). With an average percent mismatch of 10%, the reassociated duplex could ultimately be rendered almost completely acid soluble at high enough concentrations, as is suggested in the digestion of *C. cohnii* DNA in Figure 4 (○). Since the slow phase of the digestion of the reassociated DNAs in Figure 4 is initially linear, it is possible to subtract out the effect of this reaction from the more rapid phase by a linear extrapolation to zero

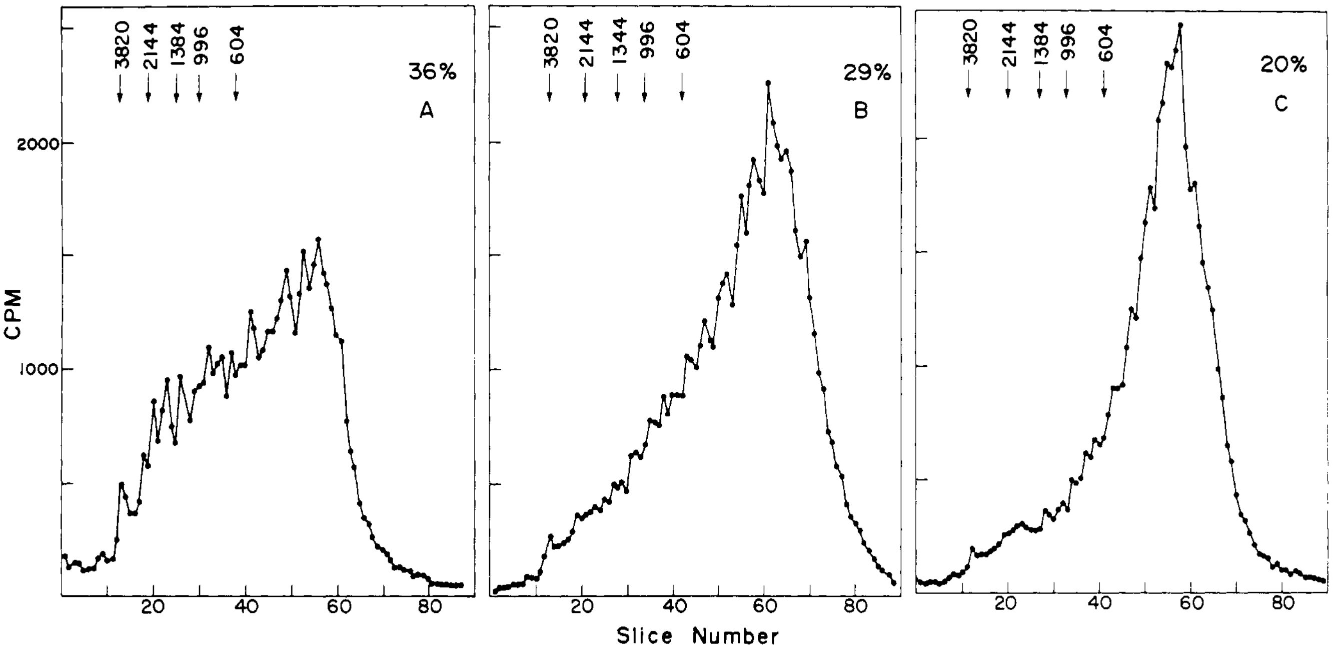

FIGURE 5: Gel electrophoretic sizing of $S_1$ nuclease digested $C_0t$ = 100 mol $L^{-1}$ s *C. cohnii* reassociated DNA fragments. For each digest, the percentage of the DNA nucleotides excluded from Sephadex G-100 and the lengths (in nucleotide pairs) and positions of the restriction fragments included as size markers are displayed across the top of each panel. For each sizing, 1.5% agarose gels and T7/*Hpa*I fragments were used. [DNA] in the digests was 125 μg/mL, and [$S_1$ nuclease] was 1.0 μL/mL in (A), 5.8 μL/mL in (B), and 11.6 μL/mL in (C).

enzyme concentration. The resulting ordinate value is the fraction of the DNA in the sample in reassociated duplex. For the multiple determinations of this quantity on *C. cohnii* DNA, we obtained a value of 38 ± 2%. This is in good agreement with the estimate of this same quantity provided by hyperchromicity measurements. On the basis of these two results, the ordinate intercept in the HA binding plot shown in Figure 2B was given the value of 0.38.

*Length Measurements on the $S_1$ Nuclease Resistant Reassociated Duplex.* The length distribution of the duplex molecules remaining after an $S_1$ nuclease digestion of long reassociated fragments is commonly used to estimate the length distribution of repeated sequences. In recent applications of this method, the effect of increasing the $S_1$ nuclease concentration on the resulting length distribution has also been examined (Britten et al., 1976; Kiper & Herzfeld, 1978). Our results using this approach are presented in Figure 5. Panels A, B, and C are agarose gel electrophoretic sizings of the $S_1$ nuclease treated *C. cohnii* DNA described in Figure 4 digested to 36, 29, and 20% exclusion from Sephadex G-100, respectively. As the digestion proceeds, the length distribution of the product becomes narrower and its average value decreases due to the loss of fragments in size classes larger than 500 nucleotide pairs. These are the expected results if the slow phase in the $S_1$ nuclease digestion (Figure 4) involves cleavage at sites of mismatch in the reassociated duplexes. The digest shown in panel A of Figure 5 is close to the best estimate that we can obtain, using this approach, for the actual length distribution of the repeated sequences, since this is the earliest point in the digestion where most of the long single-stranded regions have been adequately digested to become physically separable from the remaining duplex. We cannot expect, however, even with this limited digestion, that the longer reassociated duplexes in the sample have not been significantly reduced in size. In addition, the estimate of repeat length is further biased by the fragment length chosen for the initial reassociation step, so that the results in Figure 5A provide only a rough estimate of the actual length distribution of the repeated sequence elements.

From the data in Figure 5A we have calculated number and mass length histograms with a linear fragment length abscissa (not presented). In computing these histograms we have not included fragments under 50 nucleotide pairs in length because duplexes of this length could be seriously affected by the "nibbling" reaction well-known for $S_1$ nuclease (Shenk et al., 1975) and demonstrated for our preparation (data not shown). Both calculated distributions suggest a preponderance of short repeated sequence elements in the *C. cohnii* genome. In fact, the number distribution indicates that 80% of the repeats are shorter than 300 nucleotide pairs with an overall average value of ~200 base pairs. Since the agarose gels are not very accurate in the mobility range where most of the DNA migrates, the same digest shown in Figure 5A was sized on an 8% polyacrylamide gel and the corresponding mass and number distributions were computed. The results are in quantitative agreement with those just given, placing >75% of the duplex under 300 nucleotide pairs in length.

*Electron Microscopic Measurements of Repeated Sequence Lengths.* In view of the uncertainties involved in the interpretation of the $S_1$ nuclease results on repeated sequence lengths, we have attempted to obtain this information by using the more direct approach of electron microscopic analysis of the reassociation products formed on long strands (Manning et al., 1975; Chamberlin et al., 1975). Before reassociating samples for this analysis, we removed the zero-time binding fraction from the DNA on HA. At an estimated $C_0t$ of 3 × $10^{-3}$ mol $L^{-1}$ s, 26% of fragments with $\bar{L}_w$ = 3500 nucleotides were bound, a value close to that expected from Figure 1. After passage on HA, the $\bar{L}_w$ was reduced to 3000 nucleotides. These fragments were then reassociated to different extents, and samples were prepared for electron microscopy. Examination of samples reassociated to $C_0t$ values greater than 10 mol $L^{-1}$ s revealed the occurrence of the great majority of the strands in large network structures. A portion of a typical network is presented in Figure 6 (structure 1). Although the occurrence of networks has certain implications for sequence arrangement in *C. cohnii* (see Discussion), these structures cannot be used to measure reassociated duplex lengths because

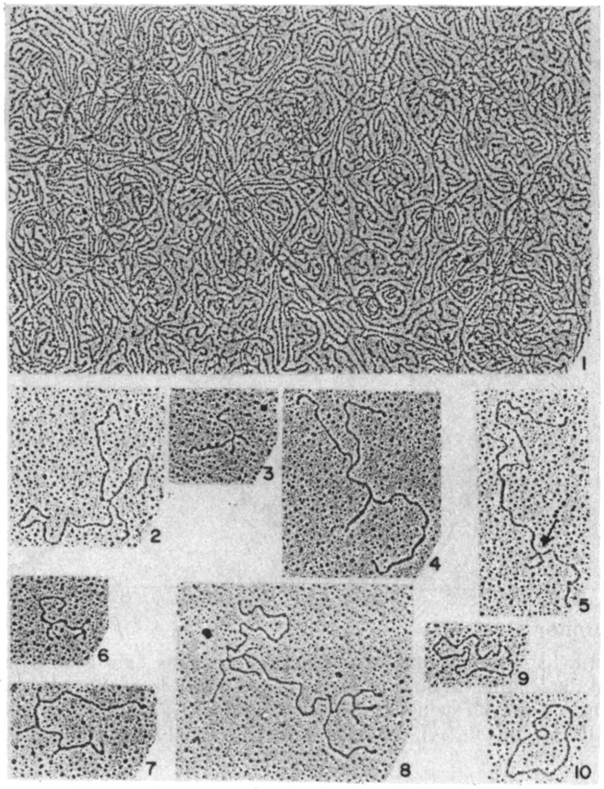

FIGURE 6: Examples of *C. cohnii* reassociation products formed on long strands of $\bar{L}_w$ = 3000 nucleotides: (1) a portion of a large network formed at $C_0t$ = 8 mol $L^{-1}$ s; (2–8) examples of structures in the multi-tailed category formed at $C_0t$ = 2 mol $L^{-1}$ s which were scored for duplex lengths. Structures 2–5 contain one measurable duplex each, structures 6 and 7 each contain two, and structure 8 contains four. Structures 9 and 10 are single- and double-stranded $\phi$X174 DNA molecules, respectively, included as length standards.

the strand density is too high to enable one to distinguish between short duplex regions and simple crossovers of unreacted strands. As a result, it was necessary to examine less reacted samples for simpler reassociation products.

A $C_0t$ value of 2 mol $L^{-1}$ s was selected for most of our measurements, where the measurable structures make up ~15% of the DNA in the spreads. Examples of these are shown in Figure 6 and include structures with one to four measurable duplex regions, each terminated by four single strands. Duplex regions terminated by less than four strands were not scored. The length measurements made on these simple structures are presented as a histogram in Figure 7A. The average value (arrow) is 250 nucleotide pairs, and 80% of the lengths are under 300 nucleotide pairs. These values are in good agreement with the number distributions calculated from the gel electrophoresis data. At $C_0t$ = 2 mol $L^{-1}$ s, although 85% of the strands can be expected to have some duplex structure already formed, only ~70% of the total repeated sequence is expected to have reacted (Figure 1). Thus, it is possible that a minor fraction of the repeated sequence length distribution has been excluded in this sample.

Since the majority of the measured duplexes are in the two smallest length classes, we were also concerned that this result could be an artifact of nonspecific crossovers of unreacted strands. We have attempted to eliminate this possibility in two ways. First, we have measured the crossover frequency of denatured fragments at known strand concentrations high enough to generate a significant number of crossovers. The number of expected crossovers in the reassociated DNA spreads from which our measurements were taken was calculated from the crossover frequency for denatured fragments and the linear molecule concentration in the scored pictures of the reassociated DNA, assuming that the number of crossovers is proportional to the square of the linear molecule concentration. Linear molecule concentrations were calculated as the average total contour length of the linears per picture. The result of this calculation is that less than 5% of the duplex lengths plotted in Figure 7A are expected to be due to crossovers.

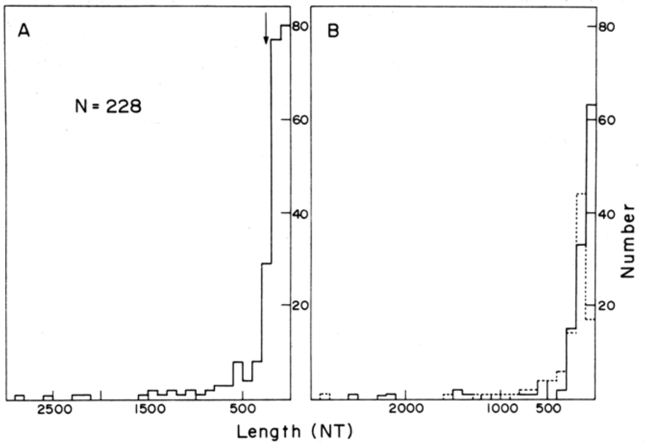

FIGURE 7: Length histograms of duplex regions on simple multi-tailed reassociation products of the kind shown in Figure 6. (A) Complete histogram of all 228 scored duplex regions. The arrow marks the average value (see text). (B) Two superimposed histograms which sum to give the histogram in panel A: (—) 99 measurements taken from spreads with 18 200 nucleotides of linear molecules per picture; (---) 129 measurements taken from spreads with 8300 nucleotides of linear molecules per picture.

The second approach we have taken is to demonstrate that the fraction of the measured lengths less than 200 nucleotide pairs is independent of the linear molecule concentration in the spreads. Half of the measurements in Figure 7A were made from spreads in which the linear concentration was roughly twice its value in the spreads from which the remaining measurements derived. If crossovers are occurring frequently, there should be a significantly higher fraction of lengths in the 0–200 nucleotide pair classes at the higher strand concentration. That this is not so is shown in Figure 7B where the measurements at the two strand concentrations are plotted separately. The fractions in the 0–200 nucleotide pair classes are 62 and 74% for the high and low concentration spreads, respectively.[2] In view of these results, we feel confident in dismissing strand crossovers as the explanation for the majority classes of observed duplex lengths.

Even at $C_0t$ = 2 mol $L^{-1}$ s, the simple structures from which measurements were made are a minority fraction of the observed structures with networks making up almost two-thirds of the DNA in the sample. As a result, we were concerned that they might represent a nontypical subclass of the repeated sequences in the genome. In an attempt to dismiss this objection, we have quantitated the occurrence of the various classes of molecules as a function of $C_0t$. These results are presented in Table I. The classes of simple structures were devised on the basis of the number of ends present, such that linears are called "two-tailed" structures. The "loop" category

[2] The fractions of duplex lengths in the 0–100 and 100–200 nucleotide pair classes shown in Figure 7B are significantly different for the spreads at the two different strand concentrations. This is probably because these lengths are at the limit of resolution of the technique and therefore are affected by small differences in the extent of shadowing. Thus, in judging the effect of strand concentration on the length distributions, we have considered these two classes combined as a single length class from 0–200 nucleotide pairs.

Table I: Quantitation of the Mass Fraction of DNA in Various Structures Appearing in Electron Micrographs of Spreads of DNA Reassociated to Three Different Values of $C_0t$[a]

| $C_0t$[b] | mass fraction in class of obsd structures | | | | | |
|---|---|---|---|---|---|---|
| (mol $L^{-1}$ s) | two-tailed | three-tailed | multi-tailed | loops | net-works | uninter-pretable |
| 0.5 | 0.467 | 0.083 | 0.170 | 0.048 | 0.300 | 0.045 |
| 2.0 | 0.211 | 0.047 | 0.124 | 0.028 | 0.637 | 0.067 |
| 8.0 | 0.168 | 0.030 | 0.053 | 0.005 | 0.761 | 0.044 |

[a] All structures were scored as if totally single stranded. This was necessary because the structure of the networks was too complicated to determine regions of base pairing unambiguously. The probable effect of this approximation is to underestimate the network class by a small amount, since these structures are expected to be the most extensively base-paired. [b] Calculation of the $C_0t$ values for the reassociated samples examined in the electron microscope was done by assuming that the zero-time binding fraction stripped from the DNA prior to reassociation is representative of the sequence content of the entire genome (Perlman et al., 1976).

will be discussed below, but for the present it can be considered as another class of simple reassociation products. At every $C_0t$ value, a fairly constant fraction of the DNA (5%) was involved in structures composed of a small number of strands which could not be interpreted as to the exact number of strands present or the arrangement of the base-paired regions. In many cases, this appeared to be due to the imperfect spreading of a normally interpretable structure falling in the "multi-tailed" category, i.e., containing ten or fewer visible ends with one to four duplex regions. However, it is possible that this was not the case and that a more complicated bonding arrangement was responsible for this class of structure. The mass fraction of the total DNA in structures of each category was determined by measuring the total contour length of the strands involved in a large number of pictures (see Table I for details). The total strand length traced at each $C_0t$ value was $3 \times 10^6$ nucleotides.

In Table I it can be seen that as $C_0t$ increases from 0.5 to 8.0 mol $L^{-1}$ s, the fraction in the two-tailed category (linears) decreased substantially, as expected from the reassociation kinetics (Figure 1). The fractions of all three of the other interpretable simple-structure classes ("three-tailed", multi-tailed, and loop) also decrease in this $C_0t$ range, by a factor of ~3, while the fraction of the DNA in large networks increases substantially. These trends are consistent with a precursor–product relationship between the simple structures and the networks. (Inclusion of the "uninterpretable" class in the multi-tailed category, where it may actually belong, lessens the magnitude of the decrease observed for the latter as $C_0t$ increases but does not invalidate the conclusion about its precursor status. In addition, the method used for measuring the fraction of DNA in the various structures underestimates somewhat the "network" class, having an effect on the above-mentioned conclusion opposite to that of the uninterpretable structures.) We conclude that these data argue in favor of the length distribution derived from the simple multi-tailed reassociation products being representative of the majority of the repeated sequences in the genome that are reassociated by $C_0t$ = 2 mol $L^{-1}$ s.

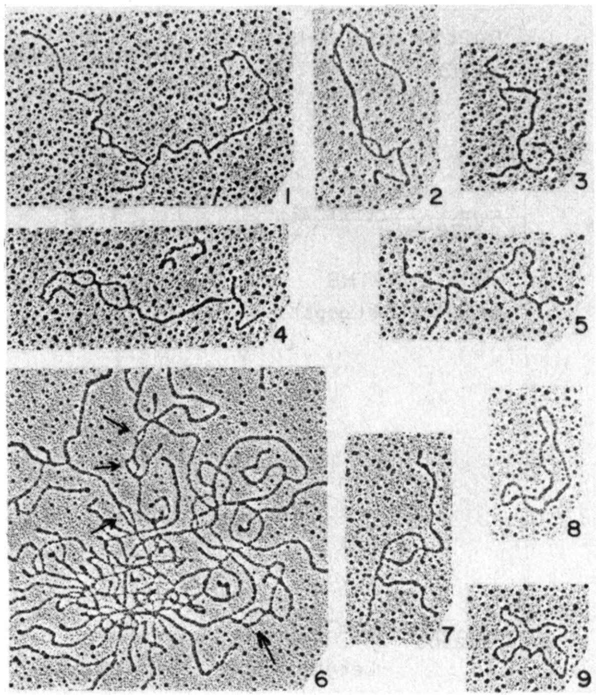

FIGURE 8: Examples of loop heteroduplexes observed in $C_0t$ = 2 mol $L^{-1}$ s reassociated *C. cohnii* DNA. Structures 1–4 are simple loop structures with very similar loop strand lengths. Structure 5 has significantly dissimilar loop strand lengths. Structure 6 is a portion of a network with several visible loops. Structure 7 is ambiguous. It could be interpreted as either a four-tailed structure of the type shown in Figure 6 (structures 2–4) or a loop structure with one of the loop duplexes only 0–200 nucleotide pairs in length. We have chosen to interpret structures of this kind as the former, since when interpreted as loop structures they show a much lower correlation between the lengths of the two loop-forming strands than unambiguous loops; 43% have strand lengths which are significantly different compared to 14% for unambiguous loop structures (see text). Structures 8 and 9 are length standards as in Figure 6.

The final problem that must be addressed in any study of this kind is that of fragment length bias. Is it possible that the length distribution in Figure 7A reflects the fragment length used in this study more than the actual repeat length distribution? We have followed Chamberlin et al. (1975) in determining the relationship between the fraction of the measured duplex that is short (<300 nucleotide pairs) and the average length of the strands involved in forming those duplexes. We have calculated that for fragment lengths between 1000 and 3500 nucleotides, this fraction is constant at ~85% and then decreases to 60% between 3500 and 4500 nucleotides. This indicates that a slight bias may occur against the longer repeated sequence classes in our measurements of the simple reassociation products. The seriousness of this bias in estimating the true repeated sequence length distribution is difficult to assess, since there are other effects which may introduce a bias in our measurements against short repeats. First, since it is expected that longer repeats will react faster than shorter repeats on fragments of the same length (due to the presence of more nucleation sites), by scoring an incomplete reassociation reaction, we may introduce a bias against shorter repeats. A second source of bias is the higher probability of a strand containing all short repeats to be involved in network formation than a strand containing even one long repeat, especially when reassociation is incomplete. This is because there will be, on average, more repeated sequence elements on the former kind of strand than the latter and thus greater opportunity for multiple reactions leading to network formation.

*Characterization of Loop Heteroduplex Structures.* In the course of making the measurements just described on the simple reassociation products, we observed an unusual type of heteroduplex that occurred as a significant fraction of the simple structures. Examples of these structures are shown in Figure 8. They appear as alternating duplex and single-stranded looped regions. As many as four loops have been observed on a single molecule. These structures are also visible in networks, as indicated in structure 6 in Figure 8. We have

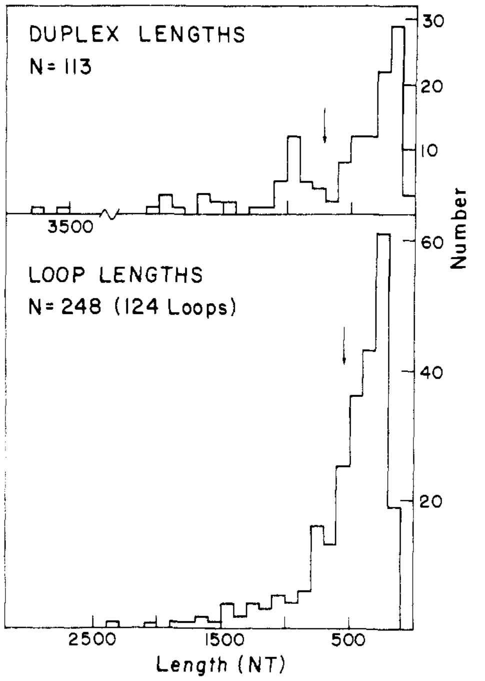

FIGURE 9: Length histograms of the duplexes and loop strands on loop structures. For each loop, the loop strands were measured separately. The arrows mark the average values (see text).

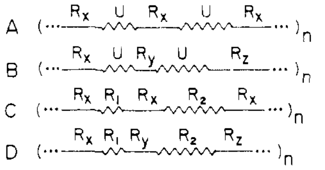

FIGURE 10: Schematic representation of various models for the sequence organization of the loop-forming repeated sequences. $R_i$ is a repeated sequence element in the *i*th family; U is a unique sequence element.

compiled the contour lengths of the duplexes and loops in simple (i.e., nonnetwork) loop structures, and the results are presented in Figure 9. It appears that the two classes of sequences are distributed differently. The duplexes are in general larger ($\bar{L}_n$ = 720 nucleotide pairs) than the loop-forming sequences (each side of the loop traced separately, $\bar{L}_n$ = 540 nucleotides). In addition, the duplexes may have a bimodal distribution. In most cases, it appeared that the two single strands forming each loop were of the same length. However, a significant fraction of the scored loops were exceptions to this rule. An example is structure 5 in Figure 8. Of the 124 loops scored, 72% have strand lengths within one standard deviation of each other. Only 14% of the loops have strand lengths different from one another at a significance level of 5% ($\alpha$ = 0.05); 8% differ at a significance level of 1% ($\alpha$ = 0.01).[3]

There are a number of different models that can be proposed to explain this class of heteroduplexes. Several of these are shown in Figure 10. In models A and B it is assumed that the loop sequences are unique DNA (U), while in C and D they are repeated sequences (R) from different families than the repeated elements which form the duplex regions. In models B and D, we depict a periodic arrangement of repeats from different families, wherein repeats from family X are often adjacent to repeats from family Y, which in turn frequently reside next to repeats of family Z. It is possible to eliminate models A and C by comparing the lengths of the adjacent duplex regions in structures containing several such regions, each terminated by two pairs of single strands. We found that 18 of the 25 observed structures of this type had adjacent duplex lengths that differed at a significance level of 1%. It is not possible to distinguish convincingly between models B and D with the data presented here. One point in favor of model B is that the length distribution of the loop-forming sequences is consistent with that expected for the interspersed unique DNA sequences based on the binding curve in Figure 2B which predicts an average value of 450–700 nucleotides (see Discussion) and the majority of the lengths less than 1000 nucleotides. Regardless of the nature of the loop-forming sequences, the repeated elements in this form of sequence arrangement do appear to be distributed over a broader range of lengths than the more simply arranged repeats shown in Figure 6. It is difficult to estimate the fraction of the repeat population of *C. cohnii* made up by these two classes, since they are expected to differ in their involvement in network structures, which contain most of the simple repeated elements in the samples we have studied.

## Discussion

*Inverted–Repeated Sequences in C. cohnii DNA.* The reassociation of *C. cohnii* tracers at $C_0t$ values of $<3 \times 10^{-5}$ mol $L^{-1}$ s (Figure 2A) is evidence for inverted repeats in the *C. cohnii* genome. In fact, recent electron microscopic studies on the $C_0t = 10^{-6}$ mol $L^{-1}$ s fraction in *C. cohnii* conducted by Yen & Rae (1978a,b) reveal the frequent occurrence of molecules with the hairpin structure expected from the "foldback" reassociation of inverted repeats. These workers have measured the lengths of duplex regions on simple "looped" and "unlooped" hairpins in *C. cohnii* and have found a distribution similar to that presented in Figure 7A for the simple noninverted repeats, with >70% of the duplex lengths between 100 and 300 nucleotide pairs (hairpins with duplex lengths less than 100 nucleotide pairs could not be scored). They also detected a minor class of unusual hairpin structures in which the duplex stems are interrupted by single-stranded loops, as occurs in the structures shown in Figure 8. As many as nine loops could be observed on a single hairpin structure. In addition, the nonhomologous strands forming each loop were often found to be very similar in length. The length distributions of the loop strands and the individual duplexes present on these hairpins are also very similar to those presented in Figure 9 with average values for the corresponding distributions within 10% of one another (C. Yen, personal communication). These similarities among the foldback and the simple repeat populations suggest that the two are related, with the former constituting a subset of the latter. Although this correspondence has also been observed for human DNA (Deininger & Schmid, 1976), it is not a general feature of the eucaryotic genome (Angerer et al., 1975; Pearson et al., 1978; Walbot & Dure, 1976).

The length dependence of the zero-time HA binding shown in Figure 2A indicates the interspersion of inverted–repeated sequences among other classes of DNA sequences. It is possible to propose a model of the organization of inverted repeats in the *C. cohnii* genome involving two classes with

[3] Determination of the significance level of the difference in loop strand lengths was carried out by computing $|L_1 - L_2|/s_D$, where $L_1$ and $L_2$ are the two loop strand lengths and $s_D$ is the expected standard deviation on the difference between two electron microscopic observations of the same single-stranded DNA molecule. $s_D$ can be shown to equal $2^{1/2}s$, where $s$ is the standard deviation on electron microscopic measurements of single-stranded molecules and is given by $s = 3.67L^{1/2}$ ($L$ in nucleotides; Davis et al., 1971). The critical values of $|L_1 - L_2|/s_D$ for deviations significant at the 5 and 1% levels are 1.96 and 3.27, respectively.

different average spacings that is consistent with all the available data. However, the HA binding data are not sufficient to warrant a detailed discussion here.

*General Repeated Sequence Organization in the C. cohnii Genome.* If certain assumptions are made, it is possible to combine the HA binding data of Figure 2B with our estimates of the total fraction of the genome in repeated sequences and the length of the interspersed repeats to obtain a model for the organization of the *C. cohnii* genome. Schmid & Deininger (1975) have presented equations for the HA binding curve of an idealized genome containing repeats of uniform length interspersed with unique sequence elements of two different lengths. The occurrence of tandemly arranged repeated sequences was also taken into account. We have analyzed our binding data by using the equations of Schmid & Deininger (1975). In doing so, we have fixed the length of the interspersed repeats at 250 nucleotide pairs. The best curve obtained by manual fitting (root mean square deviation of 0.036) is drawn in Figure 2B. The parameters of genome organization corresponding to it are 600 nucleotides for the length of the short-period interspersed unique elements, 0.50 for the mass fraction of the genome arranged in short-period interspersion, and 0.42 for the mass fraction of the genome in repeated sequences, 65% of which is tandemly arranged. This leaves 13–23% of the genome to be arranged in long-period interspersion.

We have derived a set of equations that apply to the same genome model and are formally equivalent to those of Schmid & Deininger (1975) but express the HA binding as a function of the *number* fraction of the interspersed unique elements of a given length instead of the mass fraction of the genome containing such elements (for derivation of these equations, see paragraph at end of paper regarding supplementary material). Knowledge of the number fractions of the various unique classes makes the description of sequence arrangement more complete. When applied to the curve in Figure 2B, these equations yield 0.95 as the fraction of the interspersed unique elements in short-period interspersion, leaving 5% in the long-period class with a length of >4000 nucleotides.

We have also obtained the equations for a more complicated genome model, namely, the occurrence of a uniform distribution of unique lengths in the short-period class (also presented in the supplementary material). The best computer-determined fit of the binding data to these equations has a 30% lower root mean square deviation than the curve shown in Figure 2B. However, the values of the accessible parameters are similar to those given above, with 96% of the interspersed unique elements in the short-period class, 37% of the genome mass in repeated sequence, half of which is tandemly arranged, and an average length for the short-period interspersed unique elements of 450–850 nucleotides. The scatter in the data does not permit an accurate determination of the lower limit of the short-period unique length distribution. The upper limit lies between 900 and 1200 nucleotides.

*Network Formation in C. cohnii DNA Reassociation.* The extensive network formation observed in the electron microscope for 3000-nucleotide fragments of *C. cohnii* DNA reassociated to $C_0t$ values as low as 10 mol L$^{-1}$ s is entirely consistent with the above model of DNA sequence organization. For fragments of this length, a large fraction are expected to have 3 to 4 repeats per strand. If, in general, the multiple repeats on the same strand belong to different families and the ordering of repeats from different families is not periodic, then, after reassociation of more than 1/2 to 2/3 of the repeats arranged in short-period interspersion, highly branched structures should occur (Flory, 1953). Networks comparable in dimension to those described here have also been observed in low $C_0t$ reassociation of wheat DNA using fragments of only 600–700 nucleotides in length (Flavell & Smith, 1977). This is consistent with the sequence organization in wheat wherein >50% of the genome is made up of short repeated sequences interspersed among one another and among short single-copy sequences (Flavell & Smith, 1976; Rimpau et al., 1978). On the basis of the sequence arrangement reported for the xenopus genome (Davidson et al., 1973), which is very similar to that described here for *C. cohnii*, we would also expect large network structures to occur among the low $C_0t$ reassociation products formed on long xenopus DNA fragments. However, these structures were not reported in the electron microscopic study on xenopus DNA of Chamberlin et al. (1975). We attribute this to the preselection of reassociated duplex on HA prior to the electron microscopic analysis carried out by these workers. Large networks are not expected to elute from HA with a PB elution (Thompson, 1976).

*$S_1$ Nuclease Sensitivity of Reassociated Repeated Sequences.* We have interpreted the $S_1$ nuclease digestion kinetics of reassociated *C. cohnii* DNA to be the combined result of the digestion of small single-stranded mismatch loops within duplex regions and of fully single-stranded regions. The digestion kinetics of reassociated calf DNA, also shown in Figure 6, indicate that this property is not peculiar to *C. cohnii.* There are recent reports on reassociated repeated DNA of several other eucaryotes which indicate that the length distribution of $S_1$ nuclease resistant duplexes in low $C_0t$ reassociated samples is a function of the extent of digestion, as shown in Figure 5 for *C. cohnii* (sea urchin DNA, Britten et al., 1976; cotton DNA, Walbot & Dure, 1976; parsley DNA, Kiper & Herzfeld, 1978). These studies, as well as our own, indicate the importance of examining the $S_1$ nuclease digestion kinetics of reassociated repeated sequences to obtain more accurate estimates of the fraction of the genome in repeated sequences and the length distribution of the repeats.

*Biological Implications of the Sequence Organization in C. cohnii.* The model for the organization of repeated and unique sequence elements in *C. cohnii* presented above is remarkably similar to the "xenopus pattern" of sequence arrangement described for a large number of higher plants and animals. The average length of the unique sequence elements in the short-period interspersion pattern is on the short end of the observed range but is not significantly different from that observed in certain higher plant and animal genomes. The repeated elements may also be short in comparison to other eucaryotes. The best comparisons can probably be made with xenopus, human, rat, and soybean interspersed repeats which have been studied with the electron microscope (Chamberlin et al., 1975; Deininger & Schmid, 1976; Wilkes et al., 1978; Gurley et al., 1979). The bulk of the repeated sequences in these genomes appears to be almost twice as large as in *C. cohnii.* Even so, the only really novel feature of sequence organization in *C. cohnii* is the periodic juxtaposition of members of different repeat families spaced by nonhomologous sequences of roughly equal lengths.

The functions of finely interspersed repeated sequences in the eucaryotic genome have not been defined. It has been suggested (Davidson & Britten, 1979) that they serve in some capacity to regulate the transcription of coordinately expressed genes in higher animals during development. The occurrence of this pattern of sequence arrangement in the dinoflagellate, a unicellular protist, suggests at the very least that it originally evolved for some other purpose, if any. Whatever its function, the occurrence of short-period interspersion appears to be roughly correlated with large genome size (Crain et al., 1976a). The results of our studies and recent work on euglena (Rawson et al., 1979) extend this correlation to the level of unicellular algae.

If one takes the view that dinoflagellate nuclear organization is in a primitive state, it follows that short-period interspersion of short repeats among unique DNA sequences arose very early in the evolution of the eucaryotic genome. Those species lacking this class of repeated sequences would be viewed as either having lost it at some point or having lost the ability to periodically regenerate it. These would include various green algae, fungi, nematodes, and certain insects (see beginning of paper). An alternate explanation is that this form of DNA sequence organization arose independently in a wide variety of forms, including the dinoflagellates, cellular slime molds, higher plants, and many different invertebrate and vertebrate animals. This would represent a remarkable example of convergent evolution, probably involving four separate lineages that originated independently from unicellular protista (Dobzhansky et al., 1977).

Despite the procaryotic affinities of the dinoflagellate nucleus, it is possible that these organisms are not as primitive as has been suggested. They may represent instead degenerate eucaryotes which have lost the typical eucaryotic chromatin structure and perhaps as a result have evolved an unusual mechanism of mitosis. If so, the occurrence of the typical eucaryotic sequence arrangement in the dinoflagellates would have less evolutionary significance, although the absence of histones and eucaryotic chromatin in a bona fide eucaryotic line would itself be noteworthy. It may be possible to resolve the question of dinoflagellate phylogeny with molecular sequence data. Toward this end, we have recently sequenced the 5S RNA from *C. cohnii* and we are in the process of placing this sequence on an evolutionary tree containing representatives of the major procaryotic and eucaryotic groups.

## Acknowledgments

The authors are grateful to Paul Doty for the use of facilities, Mal Smart for technical assistance, Ned Komar for expert computer programming and for initiatives on curve-fitting procedures, Gail Miller, Rick Morimoto, Nora Chapman, and Kirsten Koths for their advice on electron microscopy and gifts of reagents, Jody Miller for preparation of the figures, and Doris Boger and Nancy Hinnebusch for preparation of the manuscript. Electron microscopy was done in the EM Service Facility in the Biological Laboratories at Harvard University. A.G.H. is grateful to the National Science Foundation for a predoctoral graduate fellowship.

## Supplementary Material Available

Derivation of equations for hydroxylapatite binding curves for repeated sequence reassociation (4 pages). Ordering information is given on any current masthead page.

## References

Allen, J. R., Roberts, T. M., Loeblich, A. R., III, & Klotz, L. C. (1975) *Cell 6*, 161.

Angerer, R. C., Davidson, E. H., & Britten, R. J. (1975) *Cell 6*, 29.

Arthur, R. R., & Straus, N. A. (1978) *Can. J. Biochem. 56*, 257.

Bayen, M., & Dalmon, J. (1975) *Biochim. Biophys. Acta 395*, 213.

Beauchamp, R. S., Pasternak, J., & Straus, N. A. (1979) *Biochemistry 18*, 245.

Blin, N., & Stafford, D. W. (1976) *Nucleic Acids Res. 3*, 2303.

Britten, R. J., & Smith, J. (1970) *Carnegie Inst. Washington, Yearb. 68*, 378.

Britten, R. J., Graham, D. E., & Neufeld, B. R. (1974) *Methods Enzymol. 29E*, 363.

Britten, R. J., Graham, D. E., Eden, F. C., Painchaud, D. M., & Davidson, E. H. (1976) *J. Mol. Evol. 9*, 1.

Chamberlin, M. E., Britten, R. J., & Davidson, E. H. (1975) *J. Mol. Biol. 96*, 317.

Chamberlin, M. E., Galau, G. A., Britten, R. J., & Davidson, E. H. (1978) *Nucleic Acids Res. 5*, 2073.

Crain, W. R., Davidson, E. H., & Britten, R. J. (1976a) *Chromosoma 59*, 1.

Crain, W. R., Eden, F. C., Pearson, W. R., Davidson, E. H., & Britten, R. J. (1976b) *Chromosoma 56*, 309.

Davidson, E. H., & Britten, R. J. (1979) *Science 204*, 1052.

Davidson, E. H., Hough, B. R., Amenson, C. S., & Britten, R. J. (1973) *J. Mol. Biol. 77*, 1.

Davidson, E. H., Galau, G. A., Angerer, R. C., & Britten, R. J. (1975) *Chromosoma 51*, 253.

Davis, R. W., Simon, M., & Davidson, N. (1971) *Methods Enzymol. 21D*, 413.

Deininger, P. L., & Schmid, C. W. (1976) *J. Mol. Biol. 106*, 773.

Dobzhansky, T., Ayala, F. J., Stebbins, G. L., & Valentine, J. W. (1977) *Evolution*, pp 369–396, W. H. Freeman, San Francisco, CA.

Dodgson, J. B., & Wells, R. D. (1977) *Biochemistry 16*, 2374.

Eden, F. C., & Hendrick, J. P. (1978) *Biochemistry 17*, 5838.

Epplen, J. T., Leipoldt, M., Engle, W., & Schmidtke, J. (1978) *Chromosoma 69*, 307.

Firtel, R. A., & Kindle, K. (1975) *Cell 5*, 401.

Flavell, R. B., & Smith, D. B. (1976) *Heredity 37*, 231.

Flavell, R. B., & Smith, D. B. (1977) *Nucleic Acids Res. 4*, 2429.

Flory, P. J. (1953) *Principles of Polymer Chemistry*, pp 356–361, Cornell University Press, Ithaca, NY.

Graham, D. E., Neufeld, B., Davidson, E., & Britten, R. J. (1974) *Cell 1*, 127.

Gurley, W. B., Hepburn, A. G., & Key, J. L. (1979) *Biochim. Biophys. Acta 561*, 167.

Hamkalo, B. A., & Rattner, J. B. (1977) *Chromosoma 60*, 39.

Hinnebusch, A. G., Clark, V. E., & Klotz, L. C. (1978) *Biochemistry 17*, 1521.

Howell, S. H., & Walker, L. L. (1976) *Biochim. Biophys. Acta 418*, 249.

Hudspeth, M. E. S., Timberlake, W. E., & Goldberg, R. B. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 4332.

Kiper, M., & Herzfeld, F. (1978) *Chromosoma 65*, 335.

Lauer, G. D., Roberts, T. M., & Klotz, L. C. (1977) *J. Mol. Biol. 114*, 507.

Livolant, F., & Bouligand, Y. (1978) *Chromosoma 68*, 21.

Loeblich, A. R., III (1976) *J. Protozool. 23*, 13.

Maniatis, T., Jeffrey, A., & Kleid, D. G. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 1184.

Manning, T., Schmid, C. W., & Davidson, N. (1975) *Cell 4*, 141.

Murray, M. G., Cuellar, R. E., & Thompson, W. F. (1978) *Biochemistry 17*, 5781.

Pearson, W. R., Wu, J., & Bonner, J. (1978) *Biochemistry 17*, 51.

Perlman, S., Phillips, C., & Bishop, J. O. (1976) *Cell 8*, 33.

Rae, R. M. M. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 1141.
Rawson, J. R. Y., Eckenrode, V. K., Boerma, C. L., & Curtis, S. (1979) *Biochim. Biophys. Acta 563*, 1.
Rimpau, J., Smith, D., & Flavell, R. (1978) *J. Mol. Biol. 123*, 327.
Schildkraut, C., & Lifson, S. (1965) *Biopolymers 3*, 195.
Schmid, C. W., & Deininger, P. L. (1975) *Cell 6*, 345.
Shenk, T. E., Rhodes, C., Rigby, P. W. J., & Berg, P. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 989.
Smith, D. B., & Flavell, R. B. (1977) *Biochim. Biophys. Acta 474*, 82.
Smith, M. J., Britten, R. J., & Davidson, E. H. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 4805.
Thompson, W. F. (1976) *Plant Physiol. 57*, 617.
Timberlake, W. E. (1978) *Science 202*, 973.
Tuttle, R. D., & Loeblich, A. R., III (1975) *Phycologia 14*, 1.
Tuttle, R. D., & Loeblich, A. R., III (1977) *J. Protozool. 24*, 313.
Walbot, V., & Dure, L. S., III (1976) *J. Mol. Biol. 101*, 503.
Wells, R., Royer, H., & Hollenberg, C. (1976) *Mol. Gen. Genet. 147*, 45.
Wilkes, M. M., Pearson, W. R., Wu, J., & Bonner, J. (1978) *Biochemistry 17*, 60.
Wilson, D. A., & Thomas, C. A., Jr. (1974) *J. Mol. Biol. 84*, 115.
Yen, C. S., & Rae, P. M. M. (1978a) *J. Cell Biol. 79*, 140a.
Yen, C. S., & Rae, P. M. M. (1978b) *J. Cell Biol. 79*, 141a.

# Evidence That Deoxyribonucleic Acid Sequences Flanking the Ovalbumin Gene Are Not Transcribed[†]

Sophia Y. Tsai, Dennis R. Roop, William E. Stumph, Ming-Jer Tsai, and Bert W. O'Malley*

ABSTRACT: The transcription of DNA sequences flanking the 5′ end and 3′ end of the ovalbumin gene was examined. First, various restriction endonuclease fragments corresponding to the 5′ and 3′ regions of the gene were isolated and used as hybridization probes to assay for the presence of transcripts corresponding to these different regions in the chick oviduct nuclear RNA. Very little, if any, of the transcripts corresponding to sequences flanking the 5′ and 3′ structural sequences of the ovalbumin gene was detected in the steady-state nuclear RNA. Second, RNA was pulse labeled either in isolated nuclei or in an oviduct tissue suspension system and hybridized to DNA filters containing purified fragments of various 5′- and 3′-flanking regions. Our results again demonstrated that RNA was not synthesized from the 5′- and 3′-flanking regions surrounding the gene. Taken together, these results are consistent with the postulate that flanking DNA sequences are not transcribed and that the largest RNA species detected in the nuclear RNA are the initial transcripts.

The natural ovalbumin gene, ~7.6 kb in length, is composed of eight segments of structural DNA sequences and seven intervening sequences (Dugaiczyk et al., 1978a,b, 1979; Woo et al., 1978; Garapin et al., 1978; Lai et al., 1978; Mandel et al., 1978; Gannon et al., 1979). In order to understand how the ovalbumin gene is transcribed, we have demonstrated in our initial studies that structural and intervening sequences are transcribed at similar rates in vitro (Roop et al., 1978). However, the concentration of RNA transcripts for structural sequences is 10-fold higher than that of transcripts for intervening sequences in steady-state nuclear RNA. Also, transcripts for intervening sequences are not detected in polysomal RNA (Tsai et al., 1979). These results suggested that the entire ovalbumin gene may be transcribed as a large precursor and the intervening sequences are subsequently removed and metabolized to give rise to stable mature $\mathrm{mRNA_{ov}}$.

Further studies demonstrated that transcripts of various lengths which are larger than mature $\mathrm{mRNA_{ov}}$ and contain sequences homologous to both structural and intervening sequences of the ovalbumin gene have been found in oviduct nuclear RNA (Roop et al., 1978). The largest of these molecules detected by Northern blotting techniques was 7.8 kb in length, which is very similar to the size of the ovalbumin natural gene. These high molecular weight RNAs can be pulse labeled in an oviduct tissue suspension system, and the radioactivity can be chased into mature mRNA in the presence of an excess of unlabeled nucleoside or actinomycin D (M.-J. Tsai, A. C. Ting, J. L. Nordstrom, and B. W. O'Malley, unpublished experiments). These results support our view that the high molecular weight ovalbumin RNA detected might indeed be the precursor to mature $\mathrm{mRNA_{ov}}$.

The absence of ovalbumin nuclear RNA larger than 7.8 kb in length either in steady-state or in pulse-labeled RNA suggested that the initiation and termination of the ovalbumin gene transcripts may lie close to the 5′ end (cap site) and 3′ end [poly(A) addition site] of the mRNA. Alternatively, the ovalbumin gene could be transcribed as a precursor larger than 7.8 kb and then rapidly processed to form the 7.8-kb RNA. To develop evidence which would distinguish between these possibilities, we designed experiments which would define the existence or absence of RNA transcripts corresponding to DNA sequences flanking both 5′ and 3′ ends of the ovalbumin gene in either steady-state or pulse-labeled oviduct nuclear RNA.

## Materials and Methods

### *Materials*

Oviducts were obtained from white Leghorn chicks. The chicks were implanted subcutaneously each week with a 20-mg diethylstilbestrol (DES) pellet (Sigma Chemical Co.) which provided continuous release of DES for 8 to 9 days. Restriction endonucleases were purchased from Bethesda Research Lab-

[†] From the Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030. *Received November 15, 1979.* This work was supported by grants from the National Institutes of Health (HD-08188) and the Baylor Center for Population Research and Reproductive Biology (HD-7495).

0006-2960/80/0419-1755$01.00/0