Biochimica et Biophysica Acta, 561 (1979) 167—183

BBA 99343

SEQUENCE ORGANIZATION OF THE SOYBEAN GENOME

WILLIAM B. GURLEY *, ANGUS G. HEPBURN ** and JOE L. KEY ***

Department of Botany, University of Georigia, Athens, GA 30602 (U.S.A.)

(Received May 8th, 1978)

Key words: DNA; Genome; Sequence interspersion; (Soybean)

Summary

The total complexity of one constituent soybean (*Glycine max*) genome is estimated to be $1.29 \cdot 10^9$ nucleotide pairs, as determined by analysis of the reassociation kinetics of sheared (0.47 kilobase) DNA. Single copy sequences are estimated to represent from 53 to 64% of the genome by analysis of hydroxyapatite binding of repetitive DNA as a function of fragment length. From 65 to 70% of these single copy sequences have a short period interspersion with 1.11—1.36 kilobase lengths alternating with 0.3—0.4 kilobase repetitive sequence elements. The repetitive sequences of soybean DNA are interspersed both among themselves and among single copy regions of the genome.

Introduction

The interspersion of repetitive and single copy sequences of DNA into a highly ordered pattern appears to be a general feature of genome organization in higher eukaryotes. Although repetitive DNA has been reported in many plants [1—5], very few studies have described the pattern of sequence organization. These limited studies do, however, indicate that a portion of the repetitive sequences of both monocots and dicots are organized with an alternating arrangement of repetitive and single copy DNA. These interspersed repetitive sequences vary in size from 200 to 800 nucleotides with interspersed single copy lengths ranging from 800 to 1800 nucleotides [6—9]. Since plants often contain relatively large amounts of repetitive DNA, many repetitive regions are not contiguous with single copy sequences. The organization of these repetitive sequences can either be a tandem arrangement of highly related repeating elements typical of satellite DNA [10—12] or a complex interspersion of

* Present address: Department of Plant Pathology, 1630 Linden Drive, University of Wisconsin-Madison, 53706, U.S.A.

** Present address: Department of Biochemistry, Medical School, University Walk, Bristol BS8 ITD, U.K.

*** Research supported by Public Health Service grant CA 11624 from the National Cancer Institute and contract AT 38-1(643) from the Department of Energy.

Abbreviation: PIPES, piperazine-*N*,*N*′-bis 2-ethanesulfonic acid.

unrelated families of repetitive sequences as is the case in the genomes of wheat [7] and rye [8].

In view of the postulated roles of repetitive DNA both in chromosome structure [13] and gene regulation [14,15], we have attempted to characterize the genome of soybean (*Glycine max*) with regard to complexity, sequence reiteration, organization, size distribution, and absolute amounts of various kinetic components.

Methods

DNA Isolation. DNA was extracted and purified according to the procedure of Scott and Ingle [16] with the following modifications. Commercially obtained soybean embryonic axes (Edible Soy Products, Inc.) were imbibed at 4°C for 1 h in grinding medium (0.10 M Tris-HCl (pH 8.0)/0.02 M EDTA (pH 8.0)/0.05 M NaCl). Imbibed embryos were homogenized at a tissue to final solution volume ratio of 1 : 10 with a Brinkman Polytron. The resulting homogenate was diluted 2-fold into detergent medium with a final concentration of 0.10 M Tris-HCl (pH 8.0)/0.02 M EDTA (pH 8.0)/0.50 M NaCl/10% (v/v) *n*-butanol/6% *p*-aminosalicylate (sodium salt)/1% (w/v) triisopropylnaphthalene sulfonate (sodium salt). The mixture was then extracted once (90 min) with chloroform/isoamyl alcohol (24 : 1, v/v) and twice (1 h each) with a mixture of phenol/10% (v/v) cresol/0.1% (w/v) 8-hydroxyquinoline saturated with 0.10 M Tris-HCl (pH 8.0). The aqueous phase was collected and the nucleic acids precipitated overnight at 4°C by the addition of 2 vols. of ethanol. The resulting precipitate was pelleted (10 000 rev./min) and dissolved in 0.15 M NaCl/0.015 M trisodium citrate (1XSSC) adjusted to 20 mM EDTA (pH 8.0). RNA was degraded by digestion with pancreatic RNAase A for 2 h at 50 $\mu g \cdot ml^{-1}$ followed by digestion with T_1 RNAase for 1 h at 10 units $\cdot$ ml^{-1}. The mixture was then incubated (37°C) overnight with 400 $\mu g \cdot ml^{-1}$ of predigested (1 $mg \cdot ml^{-1}$, 1 h, 37°C) pronase. Residual protein was removed by extraction with phenol mixture and oligoribonucleotides partially removed by ethanol precipitation. Oligoribonucleotides were further removed by pelleting the DNA in 1XSSC by centrifugation at 91 000 $\times g$ (r_{av}, 6.3 cm) for 18 h at 15°C. Final purification of the DNA was achieved by banding in CsCl equilibrium density gradients (initial ρ = 1.70 $g \cdot cm^{-3}$, 35 000 rev./min) for 60 h at 25°C in a Spinco Type Ti 60 rotor. Typical yields of 8—9 mg purified DNA were obtained from 50 g dry embryos. DNA yields from whole seeds or etiolated hypocotyls were approximately one tenth this amount.

[^{3}H]DNA preparation. [^{3}H]DNA was prepared from auxin-treated soybean hypocotyls. Three-day-old, etiolated soybean seedlings were sprayed with the synthetic auxin 2, 4-dichlorophenoxyacetic acid ($2.5 \cdot 10^{-4}$ M) in order to increase DNA synthesis [17]. The seedlings were cut at the base of the hypocotyl 12 h after auxin treatment and incubated in 50 $\mu Ci \cdot ml^{-1}$ [2-^{3}H]adenosine (New England Nuclear, 23.2 μCi/mmol), 50 $\mu Ci \cdot ml^{-1}$ [*Me*-^{3}H]thymidine (New England Nuclear, 51.4 μCi/mmol) and 50 $\mu g \cdot ml^{-1}$ chloramphenicol for 12 h. DNA was extracted by grinding directly in detergent medium with mortar and pestle. The remaining purification was achieved following the procedure described above for unlabeled DNA. DNA purified from the basal 1 cm of hypocotyl had a specific activity of $1.56 \cdot 10^5$ cpm $\cdot$ μg^{-1}.

Shearing and sizing of DNA. Purified DNA was reduced to a relatively uniform low molecular weight by sonication for 10 min. This procedure resulted in a fairly homogeneous size distribution with a single strand weight average of 400—500 nucleotides as analyzed using a Beckman Model E ultracentrifuge. Sedimentation coefficients were determined and the molecular weights calculated according to Studier [18].

In vivo labeled [^{3}H]DNA from soybean hypocotyls was fractionated into relatively homogeneous size classes using 5 ml isokinetic, alkaline sucrose gradients [19,20]. Fragment sizes less than $1 \cdot 10^3$ nucleotide pairs were obtained from sonicated [^{3}H]DNA fractionated on similar gradients. Linearized ColEl plasmid [^{3}H]DNA and sonicated [^{3}H]DNA sized by analytical ultracentrifugation were included as molecular weight markers.

DNA/DNA reassociation. DNA samples ranging in concentration from 5 $\mu g \cdot ml^{-1}$ to 2.5 $mg \cdot ml^{-1}$ were denatured (100°C, 5 min) in sealed glass ampules and reassociated in either 0.12 M sodium phosphate buffer, pH 6.8, at 60°C or 0.48 M sodium phosphate buffer, pH 6.8, at 70°C. Equivalent Cot values (eCot) for reassociations in 0.48 M sodium phosphate buffer were calculated using the rate correction values of Britten et al. [21]. The incubation temperatures were 24°C below the respective T_m values of soybean DNA in these two buffers and therefore fell well within the plateau of maximum reassociation rate [22]. The reassociation reactions were terminated at the appropriate times by quick cooling in an ethanol-dry ice bath. Fractionation of single- and double-strand DNA was achieved by hydroxyapatite chromatography. The hydroxyapatite (Biorad HTP) was pre-incubated at 100°C for 5—10 min in 0.12 M sodium phosphate buffer, 0.02% (w/v) sodium lauryl sulfate (BDH, specially purified) to reduce non-specific DNA retention. Reassociated DNA samples were thawed and loaded onto jacketed hydroxyapatite columns maintained at 60°C. The single- and double-strand DNA fractions were eluted at 60°C by washing with 10 bed volumes of 0.12 M and 0.48 M sodium phosphate buffer, respectively. The amount of DNA in each fraction was determined optically by first denaturing the DNA in both fractions by adjusting to 25% (w/v) NaOH and then measuring the absorbance at 260 nm. Absorbance values measured at 320 nm were subtracted from values at 260 nm to correct for light scatter by suspended particulates.

Isolation of single-copy DNA. ^{3}H-labeled hypocotyl DNA (156 000 $cpm \cdot \mu g^{-1}$ average size 0.627 kilobase) was denatured by heating to 100°C in a sealed ampule for 5 min and incubated in 0.12 M sodium phosphate buffer at 60°C to a Cot of 50. The partly reassociated DNA was then fractionated on hydroxyapatite into the single- and double-strand components. The single-strand fraction was concentrated and dialyzed into 0.48 M sodium phosphate buffer, heat denatured, and incubated at 70°C to an equivalent Cot of 50 which corresponds to a Cot of 98 for the single copy sequences since they are the predominant sequences present. After this second reassociation, the single strand fraction was again isolated on hydroxyapatite. The final yield was 15% of the input DNA.

Optical melting. DNA samples were melted in 0.12 M sodium phosphate buffer using a Beckman Acta MVI spectrophotometer equipped with an automatic sample changer and a jacketed cuvette holder. The temperature in

the cuvette was increased at a rate of 1°C per 3 min. The DNA samples were overlaid with paraffin oil and the cuvettes sealed to prevent evaporation. Absorbance at 260 nm was corrected for thermal expansion using the published values of Mandel and Marmur [23].

Electron microscopy. Reassociated soybean DNA was examined by electron microscopy using the formamide technique of Davis et al. [24]. Samples containing 0.5 μg DNA · ml^{-1} were spread in 45% (v/v) formamide, 50—100 μg cytochrome *c* · ml^{-1} (Sigma), 0.10 M Tris-HCl (pH 8.5). This represents a criterion of T_m 12.7°C assuming a 0.72°C reduction in T_m per 1% (v/v) formamide [25]. The hypophase contained 10 mM Tris-HCl (pH 8.5) and 17% (v/v) formamide. The aqueous mounting procedure [24] was used to examine native DNA. Aqueous samples containing 0.5 μg DNA · ml^{-1} in 0.5 M ammonium acetate/1 mM EDTA (pH 7.5)/0.1 mg cytochrome *c* · ml^{-1} were spread onto a 0.25 M ammonium acetate (pH 7.5) hypophase. Grids prepared by either technique were stained in 1 μM uranyl-acetate, rotary-shadowed with 80% platinum: 20% palladium and examined using a Zeiss EM9 electron microscope. The molecular weights of duplex regions were determined by comparison with plasmid ColEl DNA ($4.2 \cdot 10^6$).

S_1 nuclease digestion. Single-strand regions of reassociated DNA were removed by digestion with S_1 nuclease purified according to the method of Vogt [26] with the omission of the sulfo-Sephadex chromatography step. The reaction mixture [27] contained 150—200 μg DNA · ml^{-1}/0.15 M NaCl/5 mM piperazine-*N*-*N*'-bis 2-ethanesulfonic acid (PIPES) buffer (pH 6.7)/40 or 25 mM sodium acetate/0.1 mM $ZnSO_4$ (pH 4.4)/25 mM β-mercaptoethanol, and twice the amount of S_1 nuclease required to completely digest an equivalent amount of single-strand DNA in 10 min. The mixture was incubated at 37°C for 2 h and the reaction terminated either by the addition of an equal volume of 0.12 M phosphate buffer or by adjusting to 20 mM EDTA (pH 8.0).

Results

A. Reassociation kinetics of sheared DNA

Fig. 1 shows the reassociation kinetics of soybean DNA sheared to an average single strand length of 0.47 kilobases. The small fraction of DNA which fails to reassociate (7%) is assumed to result from the generation of very small fragments during the random shearing process and from degradation during long incubations. Analysis of the data (solid circles) by a non-linear regression computer program gave a best fit with 3 components. The results of such an analysis are summarized in Table I with the component fractions normalized to 100% reassociation. The generation of 2 second-order curves within the repetitive fraction does not necessarily imply the existence of 2 discrete repetitive sequence families, but merely indicates the presence of more than one ideal second order component.

Since the data were obtained by hydroxyapatite chromatography fractionation, any fragment containing both repeated and single copy DNA will reassociate with the kinetics of the most highly repeated sequence. Thus, the observed DNA fractions quoted in Table I are likely to be overestimates for the repeated components and an underestimate for the single copy component. With this

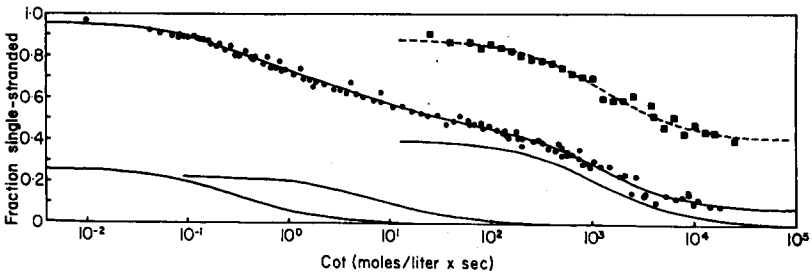

Fig. 1. Reassociation kinetics by hydroxyapatite fractionation of soybean DNA sheared to 0.47 kilobase. The solid line through data points (solid squares) was generated by a non-linear regression computer analysis for three components assuming ideal second order kinetics. The three lower curves are plots of the computer generated components. The reassociation of purified single-copy [^{3}H]DNA in the presence of a 2000-fold excess of unlabeled 0.47 kilobase total DNA is shown by the solid squares. The broken line represents the best fit non-linear regression analysis of these data (one component, RMS = 0.022) which indicates a single copy Cot 1/2 of 1263 mol · l^{-1} · s.

type of data analysis it is only possible to say that at least 43% of the total DNA consists of single copy sequences. That fraction of the DNA (0.05) which is reassociated by a Cot of $1 \cdot 10^{-2}$ mol · l^{-1} · s presumably contains very highly reiterated sequences and intrastrand duplexes.

Comparison of the observed rate constant (k_{mix}) for the single copy sequences with the rate constant and complexity of *Escherichia coli* DNA reassociated under identical conditions indicates that the kinetic complexity of one chromosomal complement (1X) is $1.29 \cdot 10^9$ nucleotide pairs, corresponding to 1.39 pg of DNA. Chemical determinations of the amount of DNA per somatic nucleus range from 5.0 to 6.5 pg DNA (Chang, H., unpublished results and ref. 28), confirming the cytological evidence of Sakai [29] that soybean is a stable tetraploid ($2n$ = 4X = 5.56 pg DNA). The close agreement between the chemically observed and the kinetically predicted $2n$ DNA values suggests a high degree of sequence homology between the constituent genomes, typical of an autotetraploid or a hybrid between very closely related species.

B. Repetitive sequences

Size distribution of inverted repeat sequences. When dilute solutions of

TABLE I

KINETIC COMPONENTS OF SOYBEAN DNA AT 0.47 KILOBASE FRAGMENT LENGTH

Component	Fraction	Normalized * fraction	Rate constant ($l \cdot mol^{-1} \cdot s^{-1}$)	Kinetic ** complexitty (nucleotide pairs)	Copies per *** constituent genome (IX)
Zero-time binding	0.04	0.05	$>1 \cdot 10^2$	—	—
Fast repeat	0.26	0.28	3.29	$9.36 \cdot 10^4$	$3.85 \cdot 10^3$
Slow repeat	0.23	0.24	$1.31 \cdot 10^{-1}$	$2.03 \cdot 10^6$	$1.53 \cdot 10^2$
Single copy	0.40	0.43	$8.55 \cdot 10^{-4}$	$5.55 \cdot 10^8$	$1.00 \cdot 10^0$

* Normalized to 100% reassocation.

** Derived from rate constants calculated for each component in a pure state.

*** Haploid genome = n = 2X.

unsheared DNA are denatured and incubated at low temperatures (0—4°C) in 0.08 M sodium phosphate buffer, intrastrand duplexes form in regions of inverted repeat or palindromic sequences [30]. Such duplexes are thought to be responsible for the fraction of 'zero-time binding' [31] observed when reassociated DNA is fractionated by hydroxyapatite chromatography at very low Cot values (i.e., less than 10^{-5}). As seen in Fig. 2A, these duplex structures may contain single-strand regions ranging from a few up to several hundred nucleotides in length as free ends and also as looped out regions between inverted repeat sequences. Digestion with the single strand specific endonuclease S_1 removes these non-base paired regions associated with the intrastrand duplex. Fig. 3 shows the size distribution of the intrastrand duplex regions of S_1 nuclease-resistant inverted repeat sequences as determined by electron microscopy. They range in length from about 50 (the limit of detection) to

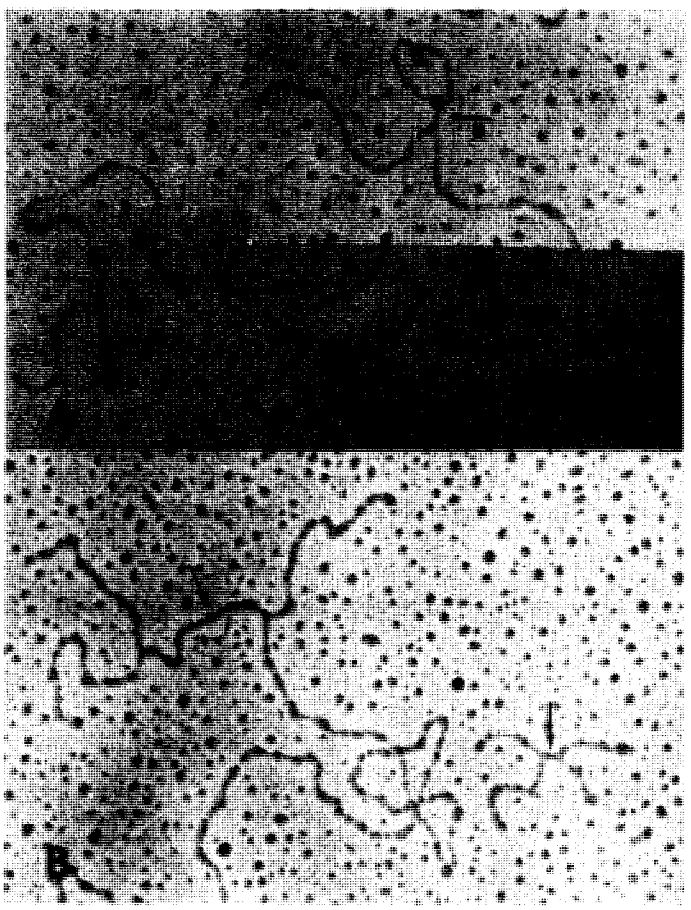

Fig. 2. Electron microscopy of reassociated DNA. (a) Foldback DNA. Intact ($6 \cdot 10^6$—$10 \cdot 10^6$ dalton) DNA was denatured, diluted into cold 0.08 M sodium phosphate buffer, and intrastrand duplexes isolated by hydroxyapatite chromatography as described by Wilson and Thomas [30]. The estimated Cot was $<10^{-5}$ mol · l^{-1} · s. (b) Cot 50 repetitive duplexes. The length of four tailed duplexes (arrows) was determined using ColE1 plasmid molecules as a double strand standard. Both Foldback and Cot 50 repetitive duplexes were examined using the formamide technique as described in Materials and Methods.

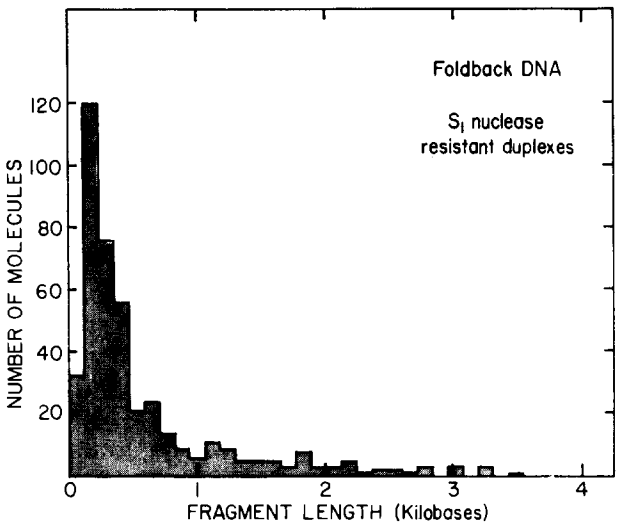

Fig. 3. Size distribution of inverted repeat duplexes. Intact ($6 \cdot 10^6$—$10 \cdot 10^6$ daltons) DNA was denatured, diluted into cold 0.08 M sodium phosphate buffer and the intrastrand duplexes isolated by hydroxyapatite (HAP) chromatography as described by Wilson and Thomas [30]. The estimated Cot is $\leqslant 1 \cdot 10^{-5}$ mol $\cdot$ l^{-1} $\cdot$ s. Samples were heated to 100°C for 3 min and cooled on ice immediately prior to S_1 nuclease digestion in order to disrupt interstrand duplexes that might have formed during sample preparation. The S_1 nuclease-resistant DNA was mounted for electron microscopy by the aqueous technique [24]. A total of 431 molecules were sized using circular ColEl plasmid DNA as a double-strand standard.

several thousand nucleotide pairs with a number average of 100—200 nucleotide pairs and a weight average of 600—700 nucleotide pairs. The size distribution of inverted repeat sequences in soybean DNA is similar to that found in other eukaryotic genomes such as HeLa and *Xenopus*.

'Zero-time binding' as a function of fragment length. As seen in Fig. 4, the fraction of DNA bound as duplex to hydroxyapatite at 'zero-time' (Cot $\leqslant 10^{-5}$

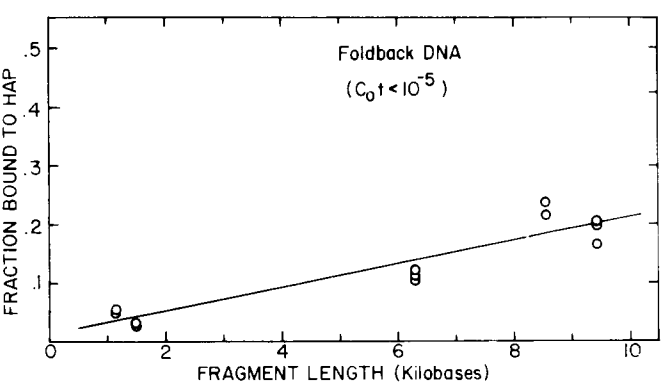

Fig. 4. 'Zero-time binding' as a function of fragment length. [^{3}H]DNA samples (1—2 ml) of various fragment lengths were heat denatured (100°C, 10 min) in sealed glass ampules at concentrations of $1 \cdot 10^{-2}$—$3 \cdot 10^{-2}$ μg $\cdot$ ml^{-1}. Denatured samples were briefly immersed in an ethanol-dry ice bath and immediately fractionated on a jacketed hydroxyapatite column at 60°C. The total elapsed time after denaturation until single strand elution was $\leqslant$80 s resulting in Cot values $\leqslant 1 \cdot 10^{-5}$ mol $\cdot$ l^{-1} $\cdot$ s.

mol · l^{-1} · s) slowly increases with increasing fragment length. The 'best fit' line generated by a linear regression analysis of the data predicts a fraction of 0.03 bound at zero fragment length. The large amount of scatter in the data and the relatively small fraction of DNA bound even with long fragment lengths renders curve fitting according to theoretical models [32,33] of inverted repeat sequences meaningless. The functions described by these models are either exponential or consist of more than one positively sloped line. The properties of these functions are such that the fraction bound (0.03) at zero fragment length represents an upper limit for the absolute fraction of the genome present in inverted repeat sequences.

Length distribution of repetitive sequence duplexes. The reassociation kinetics presented in Table I indicate that 4% of the single copy component and greater than 94% of the repeated DNA is reassociated by Cot 50. Cot 50 reassociated duplexes were analyzed by electron microscopy to determine the length distribution of repetitive sequence elements. Only those duplexes showing four single strand tails (double forked) were measured, ensuring that each measured duplex represents a complete repetitive sequence element. Examples of typical structures measured are shown in Fig. 2B. As can be seen in Fig. 5, repetitive sequences have a broad size range from those too small to distinguish unambiguously (less than 50 nucleotide pairs) to duplex lengths of several kilobases. The number average for the repetitive sequence length is 0.32 ± 0.1 kilobase and the weight average is 0.59 ± 0.11 kilobase.

Absolute amount of repetitive DNA. The hyperchromicity of reassociated

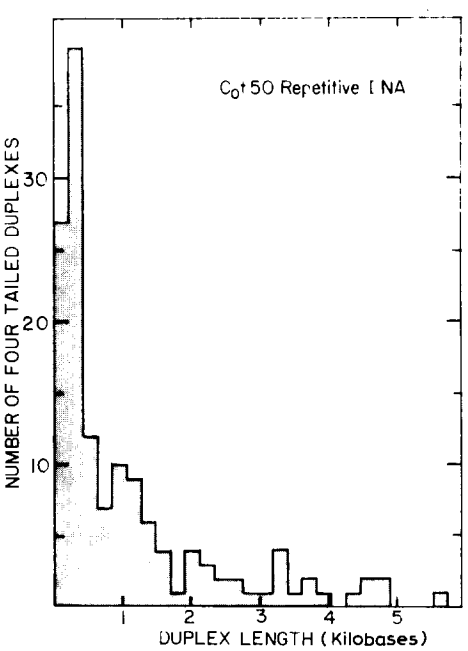

Fig. 5. Size distribution of Cot 50-repetitive duplexes. Long fragments (11 kilobases) of total DNA were reassociated to Cot 50 and examined by electron microscopy. A total of 142 double forked duplexes were measured using circular ColEl plasmid DNA as a double-strand standard.

DNA together with hydroxyapatite binding at Cot 50 can be used to determine the absolute amount of repetitive DNA present in the soybean genome. The hyperchromicity of the Cot 50 double-strand fraction is a function of the absolute amount of duplex in the fraction and, therefore, depends on the repetitive duplex length, the number of repetitive elements per fragment, and the fragment length. The extent of duplex formation is evaluated by comparing the hyperchromicity of the reassociated DNA with that of native DNA and correcting for the contribution due to the unfolding of collapsed single strand regions [27,31,33,34]. Mismatched base pairs are assumed to have no significant contribution to hyperchromicity above that of single strand collapse.

The results presented in Table II indicate that 36% of the genome is repetitive sequence DNA. It is important to note that such comparisons between native and reassociated DNA give a minimal estimate of repetitive sequence length since mismatched regions of the reassociated repetitive elements are not included. The repetitive duplex length (0.327 kilobase) obtained using 0.496 kilobase fragments is less than the weight average (0.59 kilobase) obtained from electron microscopy measurements. This result is expected when measuring hyperchromicity of fragments shorter than the weight average length of the repetitive element. The large duplex length indicated for 11 kilobase fragments is most likely due to the presence of multiple repetitive sequences per fragment.

C. Sequence organization

S_1 nuclease-resistant repetitive duplexes. The reassociation of high molecular weight enkaryotic DNA commonly results in the formation of large networks [35,36]. The size and configuration of these networks after removal of single strand regions can yield much information regarding the sequence arrangement

TABLE II

HYPERCHROMICITY OF REPETITIVE DUPLEXES

Total soybean DNA was reassociated to a Cot of 50 in 0.12 M sodium phosphate buffer at 60°C and the double-strand fraction isolated by hydroxyapatite chromatography. This duplex fraction was adjusted from 0.48 M sodium phosphate buffer to 0.12 M sodium phosphate buffer by dialysis and the hyperchromicity determined by optically monitoring thermal denaturation. The hyperchromicity of single-copy calf thymus DNA (twice reassociated to Cot 200) was determined in order to obtain a value for a single strand collapse.

Fragment length (bases)	H_{Cot50} [a]	H_{native} [a]	$H_{single\ strand\ collapse}$ [a]	Fraction [b] duplex	Duplex [c] length (bases)	Cot 50 [d] hydroxy-apatite bound (Q)	Fraction [e] genome in repetitive duplex
496	1.28	1.40	1.05	0.66	327	0.55	0.36
11 000	1.20	1.40	1.05	0.43	4730	—	—

[a] $H = A_{98}/A_{60}$.

[b] $F_{duplex} \dfrac{H_{cot50} - H_{single\ strand\ collapse}}{H_{native} - H_{single\ strand\ collapse}}$

[c] Duplex length = $F_{duplex} \times$ fragment length

[d] Normalized to 100% reassociation.

[e] $Q \times F_{duplex}$.

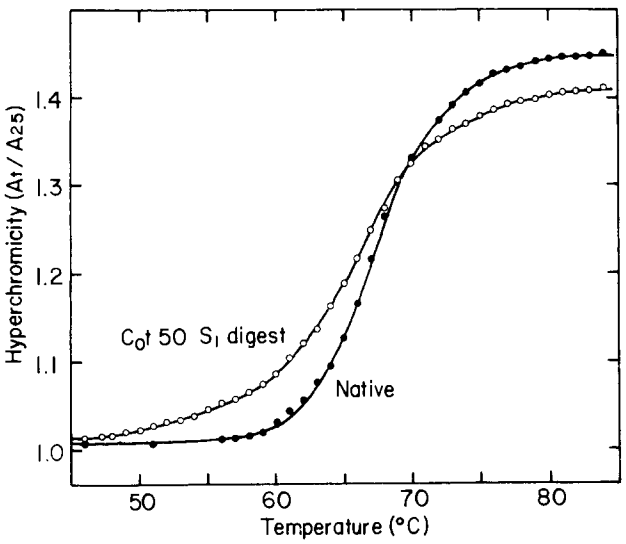

Fig. 6. Thermal denaturation of native DNA and S_1 nuclease-resistant repetitive (Cot 50) duplexes in 10.8 g/l Tris buffer/0.93 g/l Na_2 EDTA/5.5 g/l boric acid buffer.

of repetitive sequences. High molecular weight (11 kilobase) soybean DNA was denatured and allowed to reassociate to a Cot of 50. As noted earlier, at least 94% of the duplexes formed at this Cot value are from repetitive sequence DNA. The total Cot 50 reassociation mixture was digested with S_1 nuclease under conditions where the enzyme is highly specific for single strand DNA leaving double strand DNA, including mismatched regions, intact [37,38]. The hyperchromicity of the hydroxyapatite-isolated, S_1 nuclease-resistant duplexes (Fig. 6) was 93% that of native DNA, indicating nearly complete digestion of single strand regions. These repetitive duplexes were analyzed both by electron microscopy and band sedimentation.

The degree of branching present in reassociated repetitive duplexes is a function of the repetitive sequence organization. Both long repetitive units and tandem arrays of identical repetitive elements will form long unbranched duplexes upon reassociation. Although some duplexes of this type are observed with soybean DNA, many large duplex networks are also present. Examples of these highly branched networks are shown in Fig. 7. The presence of these large networks after S_1 nuclease digestion implies that the repetitive sequences found in them are organized in tandem arrays of relatively short interspersed units from unrelated sequence families. The regions between branch points represent either single repetitive units or tandem arrays of very closely related units. Branching of the reassociated duplex is the result of a variable ordering of the repetitive units present on separate single strand fragments. Short unbranched duplexes and small networks are also present in the S_1 digest implying that some repetitive sequences are contiguous with single copy or highly mismatched repetitive regions of the genome.

The results of neutral and alkaline band sedimentation of S_1 resistant repetitive duplexes are consistent with the electron microscopy observations. Analytical zone sedimentation suggests that the repetitive duplexes have a broad

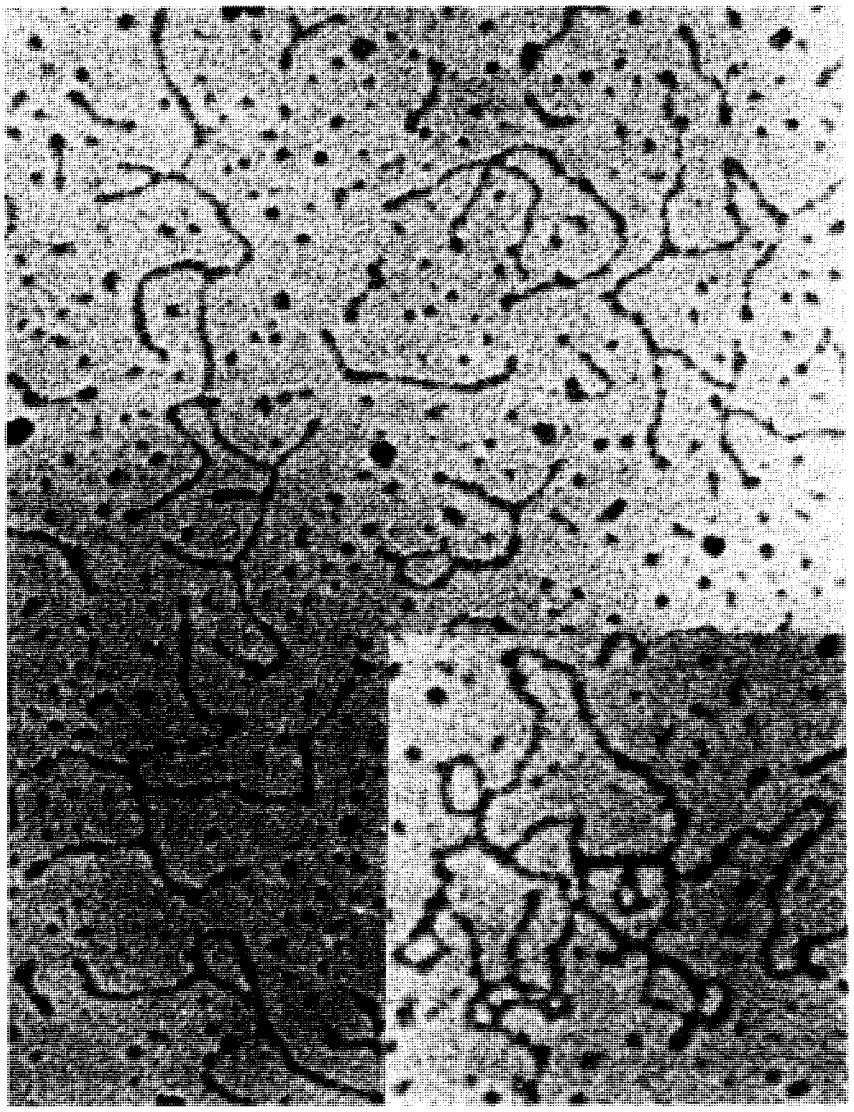

Fig. 7. Electron micrograph of S_1 nuclease-resistant repetitive duplexes. Soybean DNA was heat denatured and incubated to Cot 50 in 0.15 M NaCl/5 mM PIPES (pH 7.6) at 60°C. After cooling to 4°C, the solvent components were adjusted for S_1 nuclease digestion as described in Materials and Methods. The reaction was terminated by the addition of an equal volume of 0.12 M sodium phosphate buffer and the duplexes isolated by hydroxyapatite chromatography at 60°C. The DNA was mounted using the formamide technique. Note the presence of both simple duplexes and complex networks.

unimodal size distribution with an average sedimentation coefficient ($s^0_{20,w}$) of 12.01 in 1 M NaCl. This $s^0_{20,w}$ value corresponds to a molecular weight of $1.47 \cdot 10^6$ (2.3 kilobase pairs) using the equation of Studier [18] for native DNA. This estimate of molecular weight is only an approximation since networks may be expected to show a relationship of sedimentation to molecular weight somewhat different from that of native DNA. Sedimentation of repetitive duplexes under alkaline conditions (0.9 M NaCl, 0.1 M NaOH) gives a uni-

modal distribution of single-strand fragment size with a sedimentation coefficient of 5.97 S. The equation of Studier [18] predicts a single-strand molecular weight of $1.36 \cdot 10^5$ (0.42 kilobase), whereas that of Prunell and Bernardi [39] predicts a molecular weight of $1.10 \cdot 10^5$ (0.33 kilobase).

The presence of both simple duplexes and complex networks in the S_1 digest of Cot 50-reassociated DNA implies that the repetitive sequences of soybean DNA are organized with varying degrees of interspersion among single copy and/or highly mismatched repetitive regions. The observation that the average single-strand length (0.33—0.42 kilobase) is much shorter than the average duplex length is evidence that relatively short lengths of unrelated repetitive sequences are intermixed into long stretches of tandem repeats. Similar observations (sedimentation analysis) of S_1 resistant repetitive duplexes and single-strand lengths in wheat [7] and rye [8] genomes suggest that this type of repetitive sequence organization may be common among higher plants.

Reassociation kinetics of long tracer DNA fragments. Information regarding the organization of sequence classes can be obtained from the reassociation kinetics of tracer DNA of various lengths with uniformly sheared driver DNA. Assuming that each tracer fragment contains sequences from a single frequency class, the reassociation rate determined by hydroxyapatite chromatography is directly proportional to fragment length [31]. DNA fragments containing both single copy and repetitive sequences will have rate constants greater than those predicted solely by the effect of fragment length. As seen in Fig. 8 and summarized in Table III, the observed rates are greater than the non-interspersion model predicts for both single copy and repetitive components. These results suggest that many of the slow repetitive sequences are contiguous with fast repeats and that a portion of the single copy sequences are contiguous with a low frequency repetitive class.

The interspersion pattern of repetitive sequence elements among single copy DNA. The average spacing of repetitive sequence elements among single copy DNA was determined by reassociating trace amounts of ^{3}H-labeled DNA of

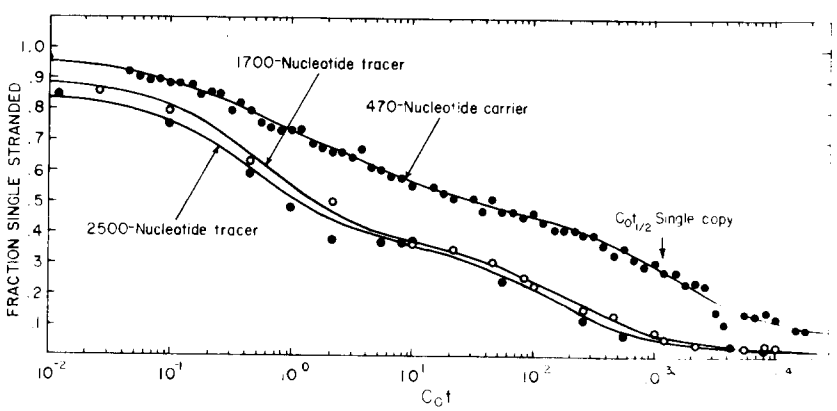

Fig. 8. Reassociation of long tracer DNA fragments. 2.5 kilobase (●) and 1.7 kilobase (○) [^{3}H]DNA fragments were reassociated (0.12 M sodium phosphate buffer, 60° C) in the presence of a 2500-fold excess of sheared total DNA (0.46 kilobase). The solid lines through the data points were generated by a non-linear regression analysis assuming two second order components.

TABLE III

REASSOCATION KINETICS OF LONG TRACER WITH SHEARED DRIVER DNA

Fragment length (kilobase)	Component	Observed k ($l \cdot mol^{-1} \cdot s^{-1}$)	Predicted k * ($l \cdot mol^{-1} \cdot s^{-1}$)	$\frac{k_{obs}}{k_{predicted}}$
1.7	single copy	$5.51 \cdot 10^{-3}$	$3.08 \cdot 10^{-3}$	1.79
1.7	repetitive	$1.88 \cdot 10^{0}$	$4.71 \cdot 10^{-1}$	3.99
2.5	single copy	$7.43 \cdot 10^{-3}$	$4.53 \cdot 10^{-3}$	1.64
2.5	repetitive	$2.08 \cdot 10^{0}$	$6.90 \cdot 10^{-1}$	3.01

* Assuming no interspersion of frequency classes: $k_{predicted} = \frac{\text{fragment length (kilobase)}}{0.47 \text{ kilobase}} \times k_{0.47 \text{ kilobase}}$.

various fragment sizes in the presence of an excess of sheared (0.47 kilobase), unlabeled DNA as driver. The reassociation was terminated at low Cot values (Cot 50 or Cot 5) such that only repetitive DNA was present in duplex structures. At very small tracer fragment sizes, little single-copy DNA will be linked to repetitive duplexes. The amount of single-copy DNA present in the 'tails' of repetitive duplexes will increase with increasing fragment size. Genomes with more than one interspersion pattern will generate a multi-slope line when hydroxyapatite binding is plotted as a function of fragment length. Analysis of these data gives an estimate of the average length of single copy sequences between repetitive elements and the absolute fraction of the genome represented in repetitive sequences. This method was first described by Davidson et al. [31] in an analysis of the interspersion of repetitive sequences in *Xenopus* DNA.

The fraction of soybean DNA containing repetitive sequences (R) as a function of fragment length (L) is shown in Fig. 9. The fraction bound to hydroxyapatite (B) was corrected [31] for 'zero-time binding' (Z) and normalized to 100% binding by the equation $R = (B - Z)/(0.93 - Z)$. The data points for each Cot value were separated into two groups and analyzed by linear regression to generate the two slopes represented as solid lines.

In view of the relatively large amount of scatter in the data, the slope of the short period interspersion line (Cot 50) in Fig. 9 was also predicted from the total fraction of repetitive DNA at a fragment length of 0.47 kilobase and the hyperchromicity of Cot 50 reassociated DNA. At a given fragment length less than the interspersed unique length the total fraction of repetitive DNA (F_r) is equivalent to B in the Cot 50 R vs. L plot. The F_r derived from the computer least squares solution (Table I) of the reassociation of sheared DNA is 0.53 (0.26 + 0.23 + 0.04 = 0.53). Point a in Fig. 9 represents the value R at 0.47 kilobase fragment length calculated from this value of F_r (0.55 = (0.53 − 0.04)/(0.93 − 0.04)). Point b in Fig. 9 was derived from the hyperchromicity of Cot 50 reassociated DNA (Table II). The intersection with the ordinate in an R vs. L plot gives an estimate of the absolute fraction of repetitive DNA at zero fragment length. An equivalent value can be obtained from the true fraction of the genome in a duplex at Cot 50. Hyperchromicity studies (Table II) indicate this fraction to be 0.36. Since the hyperchromicity of Cot 50 duplexes

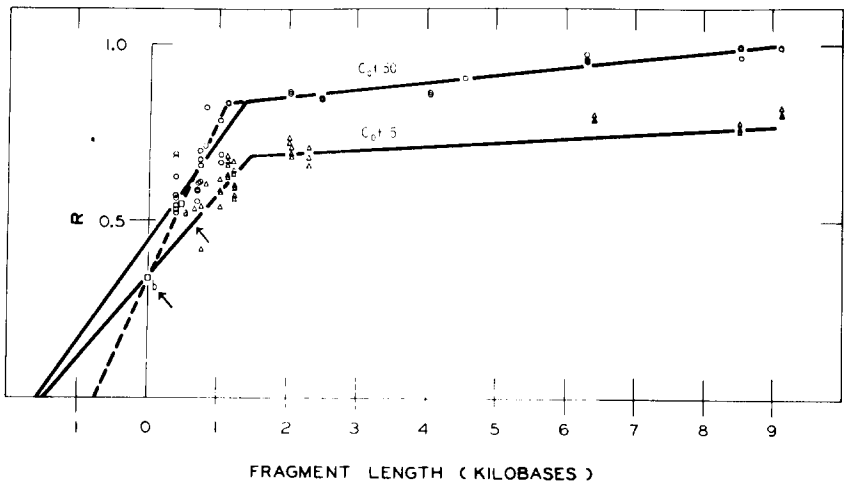

Fig. 9. Fraction of soybean DNA containing repetitive sequences as a function of fragment length. [^{3}H]-DNA fragments of various lengths were reassociated in the presence of a 5000-fold excess of unlabeled, sheared (0.40—0.50 kilobase) driver DNA to Cot 50 (○) or Cot 5 (△). The fraction bound (B) to hydroxyapatite was corrected at each fragment length for 'zero-time binding' (Z) as determined from Fig. 4 and normalized to 100% reassociation by the equation $R = (B - Z)/(0.93 - Z)$. Solid lines were generated by a linear regression analysis of the data. The slope represented by a dashed line was predicted from the fraction of the genome included in a duplex as determined by hyperchromicity studies (point b, 0.34) and the R value at Cot 50 (point a, 0.55) derived from the kinetic analysis of total DNA sheared to an average fragment length of 0.47 kilobase.

includes the contribution due to inverted repeat sequences, this value must be corrected for the fraction of the genome in an inverted repeat at zero fragment length (0.03 from Fig. 3). The value R at the ordinate point b in Fig. 9 is, therefore, predicted to be 0.34 ((0.36 − 0.03)/(0.97)). Although the predicted slope (dashed line) is derived from just two data points, it is considered to be a valid estimate due to the relatively large number of observations incorporated into each point. Based on the differences between the predicted and observed binding curves, the absolute single copy fraction is estimated to be between 0.53 and 0.64. The complexity of the single copy component is therefore $6.9 \cdot 10^8$—$8.2 \cdot 10^8$ nucleotide pairs, which is at least 160 times greater than the complexity of *E. coli* DNA.

The presence of two slopes in the Cot 50 R vs. L plot indicates that there are at least two patterns of interspersion of repetitive elements among single copy sequences. Approximately 65—70% of the single-copy sequences are organized in a short period interspersion pattern. The average distance between repetitive elements in the short pattern ranges from 1.11 to 1.36 kilobases as determined by the break points of the actual and predicted slopes. The remaining 30—35% of the single-copy sequences are organized with a longer interspersion distance averaging 9 kilobases as determined from the R vs. L plot at $R = 1.0$.

Fig. 9 also shows the fraction of DNA containing repetitive sequences as a function of fragment length at Cot 5. The slopes and break points are similar to those obtained at Cot 50 and indicate that repetitive sequences formed at Cot 5 are interspersed with a similar pattern to those formed at Cot 50.

Discussion

A fairly universal pattern of repetitive sequence organization has emerged from the study of animal genomes [40]. Most animal genomes characterized have from 50 to 80% of the DNA organized in a short period pattern with 200—600 nucleotide repetitive elements interspersed among 700—2000 nucleotide lengths of single copy sequences. The remaining repetitive elements are either interspersed with much longer lengths of single-copy sequences or are present in long stretches of tandem repeats. A major exception to this pattern is found in certain insects, including *Drosophila*, where interspersed repetitive elements are quite long (average 5.6 kilobases) with single-copy lengths averaging greater than 13 kilobases [41,42].

The organization of the soybean genome shows only a general similarity to the typical animal pattern first described in detail for *Xenopus* [31]. Repetitive DNA in soybean varies greatly in both complexity and reiteration frequency with little clear distinction between repetitive sequence clases. Although much of the repetitive DNA in soybean is not found interspersed among single-copy sequences, there appears to be a considerable amount of interspersion among repetitive classes. Evidence for the intermixing of repetitive sequence classes comes both from direct observation of complex networks in S_1 nuclease-resistant repetitive sequence networks and by analysis of long tracer DNA reassociation kinetics (Fig. 8). This type of repetitive sequence clustering is also found in other plant genomes such as wheat [7] and rye [8], and is quite different from the long tandem arrays of low complexity highly related repeated sequences present in typical animal satellites.

The short period interspersed repetitive element typical of most animal genomes has an average size of from 200 to 600 base pairs [40,43,44]. Our results using several independent methods indicate that soybean fits into this general pattern with the interspersed repetitive elements having a number average of 300—400 base pairs. Similar studies with cotton [6], wheat [7], and tobacco [9] show a similar sized (200—800 base pairs) repetitive element interspersed among single copy sequences. Although the size range of repetitive sequences appears broader in plants than in animals [6], there is a remarkable similarity in the average size of the repetitive element found interspersed among single-copy sequences in eukaryotes. The fairly universal presence of a short repetitive element in eukaryotes could be (1) the result of a common mechanism of sequence rearrangement which generates the interspersion of unrelated sequences, and/or (2) related to an optimum size required for repetitive sequence functions.

Since most structural genes are apparently encoded by single-copy DNA [45—47], the average length of a component of the single copy DNA must be no smaller than the average size of mRNA. In soybean, approximately 65—70% of the single copy sequences have an average length of 1.11—1.36 kilobases. The remaining 30—35% have an average length of 9.0 kilobases. The average length of the poly(A)-containing mRNA of soybean is about 1300 nucleotides excluding the poly(A) 'tail' which averages 100—140 nucleotides [48]. The close size correspondence of poly(A)-containing mRNA and the average length of the interspersed single copy DNA sequences is consistent with the notion

that mRNA may be transcribed from the latter. This concept is supported by the results of Davidson et al. [20] showing that 80—100% of embryonic mRNA from sea urchin is transcribed from single copy DNA adjacent to interspersed repetitive sequences. The number of short period single copy elements per constituent genome ranges from $3.4 \cdot 10^5$ to $5.2 \cdot 10^5$ in soybean. However, it appears unlikely that each of these single-copy elements represents a structural gene since the complexity of soybean poly(A) RNA is only 6 to 8% of the genome DNA complexity [49] corresponding to roughly $3.0 \cdot 10^4$ average sized mRNA molecules.

A short period interspersion pattern of single copy sequences has been shown to occur in a wide variety of plant and animal genomes separated by great evolutionary distances. The broad universality of this type of sequence arrangement suggests that (1) this pattern may represent an ancient adaptation made by eukaryotes or their progenitors, and (2) genome sequence organization may have a high selective value and, hence, a direct functional significance. The correlation between the interspersed single copy sequence length and the average size of mRNA, and the finding that most structural genes are contiguous with interspersed repetitive DNA sequences in sea urchin [20] are at least compatible with the concept that some of this selective value may be related to the regulation of gene transcription as proposed by Britten and Davidson [14,40,50].

References

1 Britten, R.J. and Kohn, D.E. (1967) Carnegie Inst. Wash. Yearbook 65, 73—88
2 Bendich, A.J. and McCarthy, B.J. (1970) Genetics 65, 545—565
3 Bendich, A.J. and McCarthy, B.J. (1970) Genetics 65, 567—573
4 Nze-Ekekang, L., Patillon, M., Schafer, A. and Kovoor, A. (1974) J. Exp. Bot. 25, 320—329
5 Flavell, R.B., Bennett, M.D., Smith, J.B. and Smith, D.B. (1974) Biochem. Genet. 12, 257—269
6 Walbot, V. and Dure, L.S., III (1976) J. Mol. Biol. 101, 503—536
7 Flavell, R.B. and Smith, D.B. (1976) Heredity 37, 231—252
8 Smith, D.B. and Flavell, R.B. (1977) Biochim. Biophys. Acta 474, 82—97
9 Zimmerman, J.L. and Goldberg, R.B. (1977) Chromosoma 59, 227—252
10 Ingle, J., Pearson, G.G. and Sinclair, J. (1973) Nat. New Biol. 242, 193—197
11 Bendich, A.J. and Taylor, W.C. (1977) Plant Physiol. 59, 604—609
12 Timmis, J.N., Deumling, B. and Ingle, J. (1975) Nature 257, 152—155
13 Walker, P.M.B. (1971) Nature 229, 306—308
14 Britten, R.J. and Davidson, E.H. (1969) Science 165, 349—357
15 Georgiev, G.P. (1969) J. Theor. Biol. 25, 473—490
16 Scott, N.S. and Ingle, J. (1973) Plant Physiol. 51, 677—684
17 Key, J.L., Lin, C.-Y., Gifford, E.M., Jr. and Dengler, R. (1966) Bot. Gaz. 127, 87—94
18 Studier, F.W. (1965) J. Mol. Biol. 11, 373—390
19 Noll, H. (1967) Nature 215, 360—363
20 Davidson, E.H., Hough, B.R., Klein, W.H. and Britten, R.J. (1975) Cell 4, 217—238
21 Britten, R.J., Graham, D.E. and Neufeld, B.R. (1971) in Methods in Enzymology, (Grossman, L. and Moldave, K., eds.), Vol. 29, pp. 413—428, Academic Press, New York
22 Bonner, T.I., Brenner, D.J., Neufeld, B.R. and Britten, R.J. (1973) J. Mol. Biol. 81, 123—135
23 Mandel, M. and Marmur, J. (1968) in Methods in Enzymology, (Grossman, L. and Moldave, K., eds.), Vol. 12, part B, pp. 195—206, Academic Press, New York
24 Davis, R.W., Simon, M. and Davidson, N. (1971) in Methods in Enzymology, (Grossman, L. and Moldave, K., eds.), Vol. 21, pp. 413—428, Academic Press, New York
25 McConaughy, B.L., Laird, C.D. and McCarthy, B.J. (1969) Biochem. 8, 3289—3295
26 Vogt, V.M. (1973) Eur. J. Biochem. 33, 192—200
27 Angerer, R.C., Davidson, E.H. and Britten, R.J. (1975) Cell 6, 29—39
28 Sparrow, A.H. and Miksche, J.P. (1961) Science 134, 282—283
29 Sakai, B. (1951) La Kromosoma 11, 425—429

30 Wilson, D.A. and Thomas, C.A., Jr. (1974) J. Mol. Biol. 84, 115—144
31 Davidson, E.H., Hough, B.R., Amenson, C.S. and Britten, R.J. (1973) J. Mol. Biol. 77, 1—23
32 Hammer, D.H. and Thomas, C.A., Jr. (1974) J. Mol. Biol. 84, 139—144
33 Schmid, C.W. and Deininger, P.L. (1975) Cell 6, 345—358
34 Graham, D.E., Neufeld, B.R., Davidson, E.H. and Britten, R.J. (1974) Cell 1, 127—137
35 Britten, R.J. and Waring, M.J. (1965) Carnegie Inst. Wash. Yearbook 64, 316—321
36 Bendich, A.J. and Bolton, E.T. (1967) Plant Physiol. 42, 959—967
37 Davidson, E.H., Graham, D.E., Neufeld, B.R., Chamberlin, M.R., Amenson, C.S., Hough, B.R. and Britten, R.J. (1973) Cold Spring Harbor Symp. Quant. Biol. 38, 295—301
38 Rice, N.R. (1974) Carnegie Inst. Wash. Yearbook 73, 1094—1098
39 Prunell, A. and Bernardi, G. (1973) J. Biol. Chem. 248, 3433—3440
40 Davidson, E.H. and Britten, R.J. (1973) Quart. Rev. Biol. 48, 565—613
41 Manning, J.E., Schmid, C.W. and Davidson, N. (1975) Cell 4, 141—155
42 Crain, W.R., Eden, F.C., Pearson, W.R., Davidson, E.H. and Britten, R.J. (1976) Chromosoma 56, 309—326
43 Davidson, E.H., Galau, G.A., Angerer, R.C. and Britten, R.J. (1975) Chromosoma 51, 253—259
44 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. and Davidson, E.H. (1975) Chromosoma 51, 225—251
45 Greenberg, J.R. and Perry, R.P. (1971) J. Cell Biol. 50, 774—786
46 Firtel, R.A., Jacobson, A. and Lodish, H.F. (1972) Nat. New Biol. 239, 225—228
47 Goldberg, R.B., Galau, G.A., Britten, R.J. and Davidson, E.H. (1973) Proc. Natl. Acad. Sci. U.S. 70, 3516—3520
48 Key, J.L. and Silflow, C. (1975) Plant Physiol. 56, 364—369
49 Silflow, C. (1977) Ph.D. Thesis, University of Georgia
50 Davidson, E.H., Klein, W.H. and Britten, R.J. (1977) Dev. Biol. 55, 69—84