The Biochemistry of Plants

A COMPREHENSIVE TREATISE

P. K. Stumpf and E. E. Conn

EDITORS-IN-CHIEF

Volume 6

Proteins and Nucleic Acids

Abraham Marcus

EDITOR

The Biochemistry of Plants

A COMPREHENSIVE TREATISE

Volume 6

P. K. Stumpf and E. E. Conn

EDITORS-IN-CHIEF

Department of Biochemistry
and Biophysics
University of California
Davis, California

Volume 1 **The Plant Cell**
N. E. Tolbert, Editor

Volume 2 **Metabolism and Respiration**
David D. Davies, Editor

Volume 3 **Carbohydrates: Structure and Function**
Jack Preiss, Editor

Volume 4 **Lipids: Structure and Function**
P. K. Stumpf, Editor

Volume 5 **Amino Acids and Derivatives**
B. J. Miflin, Editor

Volume 6 **Proteins and Nucleic Acids**
Abraham Marcus, Editor

Volume 7 **Secondary Plant Products**
E. E. Conn, Editor

Volume 8 **Photosynthesis**
M. D. Hatch and N. K. Boardman, Editors

THE BIOCHEMISTRY OF PLANTS

A COMPREHENSIVE TREATISE

Volume 6

Proteins and Nucleic Acids

Abraham Marcus, editor

The Institute for Cancer Research
Fox Chase Cancer Center
Philadelphia, Pennsylvania

1981

ACADEMIC PRESS

A Subsidiary of Harcourt Brace Jovanovich, Publishers

New York London Toronto Sydney San Francisco

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DX

Library of Congress Cataloging in Publication Data
Main entry under title:

The Biochemistry of plants.

Includes bibliographies and indexes.
CONTENTS: v. 1. The plant cell.--v. 2. Metabolism and respiration.--[etc.]--v. 6. Proteins and nucleic acids.
1. Botanical chemistry. I. Stumpf, Paul Karl, 1919–
II. Conn, Eric E.
QK861.B48 581.19'2 80–13168
ISBN 0–12–675406–3 (v. 6)

PRINTED IN THE UNITED STATES OF AMERICA

81 82 83 84 9 8 7 6 5 4 3 2 1

Contents

List of Contributors

Numbers in parentheses indicate the pages on which the authors' contributions begin.

George Bruening (571), Department of Biochemistry and Biophysics, University of California, Davis, California 95616

T. A. Dyer (111), Plant Breeding Institute, Trumpington, Cambridge CB2 2LQ, England

Marvin Edelman (249), Department of Plant Genetics, The Weizmann Institute of Science, Rehovot, Israel

S. M. Flashman* (83), Department of Genetics, North Carolina State University, Raleigh, North Carolina 27650

M. P. Gordon (531), Department of Biochemistry, University of Washington, Seattle, Washington 98195

Tom J. Guilfoyle (207), Department of Botany, Biological Science Center, St. Paul, Minnesota 55108

Brian A. Larkins (449), Department of Botany and Plant Pathology, Purdue University, West Lafayette, Indiana 47907

C. J. Leaver (111), Department of Botany, University of Edinburgh, Edinburgh EH9 3JH, Scotland

C. S. Levings III (83), Department of Genetics, North Carolina State University, Raleigh, North Carolina 27650

Halina Lis (371), Department of Biophysics, The Weizmann Institute of Science, Rehovot, Israel

Michael G. Murray (1), Department of Plant Biology, Carnegie Institute of Washington, Stanford, California 94305

* Present address: DeGuigné Technical Center, Stauffer Chemical Company, Richmond, California 94804.

Cleon W. Ross (169), Department of Botany and Plant Pathology, Colorado State University, Fort Collins, Colorado 80523

C. A. Ryan (321, 351), Department of Agricultural Chemistry, Washington State University, Pullman, Washington 99164

Nathan Sharon (371), Department of Biophysics, The Weizmann Institute of Science, Rehovot, Israel

Katherine E. Steinback* (303), Department of Botany, University of Illinois, Urbana, Illinois 61801

William F. Thompson (1), Department of Plant Biology, Carnegie Institution of Washington, Stanford, California 94305

M. Walker-Simmons (321), Department of Agricultural Chemistry, Washington State University, Pullman, Washington 99164

Donald P. Weeks (491), The Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111

* Present address: Department of Biological Sciences, Wellesley College, Wellesley, Massachusetts 02181

General Preface

In 1950, James Bonner wrote the following prophetic comments in the Preface of the first edition of his "Plant Biochemistry" published by Academic Press:

> There is much work to be done in plant biochemistry. Our understanding of many basic metabolic pathways in the higher plant is lamentably fragmentary. While the emphasis in this book is on the higher plant, it will frequently be necessary to call attention to conclusions drawn from work with microorganisms or with higher animals. Numerous problems of plant biochemistry could undoubtedly be illuminated by the closer application of the information and the techniques which have been developed by those working with other organisms . . .
>
> Certain important aspects of biochemistry have been entirely omitted from the present volume simply because of the lack of pertinent information from the domain of higher plants.

The volume had 30 chapters and a total of 490 pages. Many of the biochemical examples cited in the text were derived from studies on bacterial, fungal, and animal systems. Despite these shortcomings, the book had a profound effect on a number of young biochemists since it challenged them to enter the field of plant biochemistry and to correct "the lack of pertinent information from the domain of higher plants."

Since 1950, an explosive expansion of knowledge in biochemistry has occurred. Unfortunately, the study of plants has had a mixed reception in the biochemical community. With the exception of photosynthesis, biochemists have avoided tackling for one reason or another the incredibly interesting problems associated with plant tissues. Leading biochemical journals have frequently rejected sound manuscripts for the trivial reason that the reaction had been well described in *E. coli* and liver tissue and thus was of little interest to again describe its presence in germinating pea seeds! Federal granting agencies, the National Science Foundation excepted, have

also been reluctant to fund applications when it was indicated that the principal experimental tissue would be of plant origin despite the fact that the most prevalent illness in the world is starvation.

The second edition of "Plant Biochemistry" had a new format in 1965 when J. Bonner and J. Varner edited a multiauthored volume of 979 pages: in 1976, the third edition containing 908 pages made its appearance. A few textbooks of limited size in plant biochemistry have been published. In addition, two continuing series resulting from the annual meetings and symposia of phytochemical organizations in Europe and in North America provided the biological community with highly specialized articles on many topics of plant biochemistry. Plant biochemistry was obviously growing.

Although these publications serve a useful purpose, no multivolume series in plant biochemistry has been available to the biochemist trained and working in different fields who seeks an authoritative overview of major topics of plant biochemistry. It therefore seemed to us that the time was ripe to develop such a series. With encouragement and cooperation of Academic Press, we invited six colleagues to join us in organizing an eight volume series to be known as "The Biochemistry of Plants: A Comprehensive Treatise." Within a few months, we were able to invite over 160 authors to write authoritative chapters for these eight volumes.

Our hope is that this Treatise not only will serve as a source of current information to researchers working in plant biochemistry, but equally important will provide a mechanism for the molecular biologist who works with *E. coli* or the neurobiochemist to become better informed about the interesting and often unique problems which the plant cell provides. It is hoped, too, the senior graduate student will be inspired by one or more comments in chapters of this Treatise and will orient his future career to some aspect of this science.

Despite the fact that many subjects have been covered in this Treatise, we make no claim to have been complete in our coverage nor to have treated all subjects in equal depth. Notable is the absence of volumes on phytohormones and on mineral nutrition. These areas, which are more closely associated with the discipline of plant physiology, are treated in multivolume series in the physiology literature and/or have been the subject of specialized treatises. Other topics (e.g., alkaloids, nitrogen fixation, flavonoids, plant pigments) have been assigned single chapters even though entire volumes, sometimes appearing on an annual basis, are available.

Finally, we wish to thank all our colleagues for their enthusiastic cooperation in bringing these eight volumes so rapidly into fruition. We are grateful to Academic Press for their gentle persuasive pressures and we are indebted to Ms. Barbara Clover and Ms. Billie Gabriel for their talented assistance in this project.

P. K. Stumpf
E. E. Conn

Preface to Volume 6

Nucleic acids have been the focus of a recent upsurge of new biological insights. As a result, much current research emphasizes the understanding of the information present in the nucleic acids and the regulation of the expression of this information (selective transcription). In this volume we stress these recent developments, at the same time keeping in mind that the processes by which the nucleic acids are finally expressed as proteins (processing and translation), and the knowledge of the proteins themselves remain the fundamental fabric of the biological system. Detailed information on these components provides the molecular description of cellular phenotype. Regulation of the function of the proteins *in vivo,* although considerably elaborated *in vitro,* is still to a large part an uncharted area of biological control.

The sequence of presentation follows the lines: DNA → RNA → protein with detailed descriptions of those classes of proteins found predominantly in plants. A notable absence are those proteins, sometimes termed chromoproteins, associated with respiration and photosynthesis. These are discussed in Volumes 2 and 8 and elsewhere in this Treatise. The final chapters cover two special areas of plant biology where the expression of the nucleic acids is seen in striking relief, the formation of plant tumors and the growth and expression of plant viruses.

Throughout we have tried to be thorough in coverage with emphasis on principles, and have stressed projections of areas of future import. Needless to say much noteworthy work has, of necessity, been considered only briefly. Finally, an expression of personal gratitude is due to all of the contributors who made this volume possible.

Abraham Marcus

This page intentionally left blank

The Nuclear Genome: Structure and Function*

1

WILLIAM F. THOMPSON
MICHAEL G. MURRAY

I. DNA SEQUENCE ORGANIZATION

A. Introduction

DNA is the genetic material of all organisms except RNA viruses, and it is natural to think of DNA almost exclusively in terms of genes and their controlling elements. However, if genes and their controlling elements actually made up a large fraction of the total DNA, we would expect that the DNA content of different organisms would be directly correlated with organism complexity since the evolution of more elaborate developmental and

* This manuscript was submitted for publication in December 1978.

The Biochemistry of Plants. Vol. 6

ISBN 0-12-675406-3

physiological processes should require the evolution of more genes and/or controlling elements.

Indeed, the *minimum* DNA content for various major taxa does increase in a more-or-less regular fashion with increasing complexity (Sparrow *et al.*, 1972). These minimum values might be regarded as approximate indications of the amount of DNA essential for genes and controlling elements at each level of complexity. However, even more striking is the tremendous range of variation in DNA content encountered even among organisms with similar biological complexities. Eukaryotic animals generally contain between 1 and 5 pg of DNA per haploid genome, but the range for salamanders extends to 100 pg, and primitive vascular plants may contain as much as 300 pg (Hinegardner, 1976). More advanced vascular plants frequently have much lower genome sizes than psilopsids and ferns, although there are large differences even among the modern forms; e.g., angiosperm species exhibit a nearly 100-fold range of variation. Significant differences are seen even within more restricted taxonomic groups such as genera. One well-known example is the genus *Vicia,* in which DNA content may differ up to 7-fold even among diploid species with $2n = 14$ chromosomes (Chooi, 1971; Bennett and Smith, 1976). On the basis of extensive studies on life history features in herbaceous angiosperms with different DNA contents, Bennett (1972) has concluded that "nuclear DNA content in higher plants is not positively correlated with either the amount or the sophistication of genotypic control exercised during development."

Even among animals with genomes which are small by higher-plant standards, available evidence suggests that only 1–10% of the genome is required to account for all known functions of DNA in development. These estimates were obtained in several ways, including calculations of the amount of DNA required to code and control a "reasonable" number of genes, estimates of genetic load based on the frequency of observable lethal mutations, direct measurements of the fraction of the genome transcribed into nuclear RNA and messenger RNA, and comparisons of the rates of evolutionary change in coding sequences and total DNA sequence populations (e.g., Rosbach *et al.*, 1975; Angerer *et al.*, 1976; Britten and Davidson, 1976; Hinegardner, 1976).

The large and variable sizes of eukaryotic genomes in general, and those of plants in particular, suggest therefore that much of the genetic material in these organisms is not needed for known coding or regulatory functions in development. What, then, is all this excess DNA doing in so many genomes? How did it get there? Can it really be useless or functionally redundant, or might it serve purposes not yet perceived? The first part of this chapter reviews basic data on the overall structure of eukaryotic genomes, with particular emphasis on higher-plant DNA. This information provides a basis upon which we can begin to speculate about these questions and also brings

us face to face with the problems that must be solved to proceed beyond speculation to real understanding.

B. DNA Reassociation

Many of our current concepts of genome organization are derived from DNA reassociation experiments in which the two component strands of double-helical DNA fragments are separated and fragments bearing complementary sequences then allowed to reform base-paired structures during incubation under carefully defined conditions of salt and temperature. Much of the theory and practice of commonly used techniques are reviewed in detail by Britten *et al.* (1974).

The major steps in a typical reassociation experiment are outlined in Fig. 1. Strand separation ("melting") during the denaturation process results in an increase in absorbance, or a hyperchromic shift. Absorbance-versus-temperature curves can provide detailed information concerning the thermal stability of duplex DNA (e.g., Cuellar *et al.*, 1978). Average thermal stability is conveniently expressed as the T_m, the temperature at which 50% of the hyperchromic shift has occurred. In the case of reassociated DNA molecules containing mismatched base pairs, T_m is reduced in proportion to the extent of mismatching (Britten *et al.*, 1974).

Reassociation is initiated by rapidly cooling the sample to a predetermined appropriate temperature. Proper choice of temperature and salt concentration is important, as these factors determine the stringency of the reassociation—that is, the precision of base pairing required for formation of a stable duplex molecule. Increased stringency is obtained by raising the temperature or lowering the salt (monovalent cation) concentration. For simple DNA which reassociates precisely with little or no mispairing, the optimum rate of reassociation is achieved at temperatures in the vicinity of 25°C below the T_m of native DNA in the given buffer (Wetmur and Davidson, 1968). Salt concentration may also affect the rate of the reaction at the optimum temperature, with reassociation generally proceeding faster at higher salt concentrations (Britten *et al.*, 1974).

Three techniques are commonly used to measure the progress of reassociation. Since strand separation produces a hyperchromic shift which is reversed as base pairs reform, reassociation may be followed by monitoring the decrease in optical density (the hypochromic shift) as a function of time. Extremely precise measurements can be made this way, and the technique is nondestructive; a single sample of DNA can be monitored continuously over a long period of time. The range of concentrations which can be employed in optical experiments is limited, however, and hypochromicity measurements do not permit one to follow the behavior of different DNA samples in a mixture, as is possible when using tracers in other assay systems.

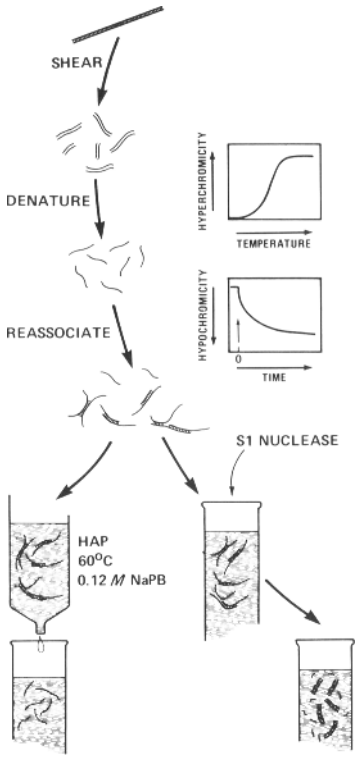

Fig. 1 Some essential steps in reassociation experiments. Starting with high-MW native DNA, mechanical shearing is used to reduce the fragment length to an appropriate value. Except when long DNA is deliberately used in interspersion experiments, it is desirable to work with the shortest possible length in order to maximize the physical separation of repeated and single-copy sequences and minimize the formation of complex multimolecular structures during reassociation. Prior to reassociation, the two strands of the helix are separated by thermal denaturation or exposure to alkali. Separation of the strands results in an increase in optical density (hyperchromicity); when this increase is monitored as a function of temperature, one can obtain useful information about the thermal stability of the helix (see text). Upon lowering the temperature, reassociation occurs as base pairs form between complementary or partially complementary strands. Since base pairing reverses the hyperchromic shift which occurred upon strand separation, the progress of reassociation can be monitored as a decrease in optical density, or hypochromic shift. Initially, reassociation results in the formation of two-stranded duplexes. However, since randomly sheared DNA fragments are used, the fragments in any given duplex will normally not be complementary over their entire length. In addition, when short repeated sequences are present, reassociation will usually produce repetitive duplexes with noncomplementary single-copy sequences attached to the duplex region. In either case, most reassociation products will contain single-stranded tails, some of which may participate in secondary reactions later in the course of the experiment. In addition to optical techniques, two other methods are commonly used to measure reassociation. As illustrated, hydroxyapatite (HAP) chromatography separates duplexes and their associated single-stranded tails from completely single-stranded fragments, while S1 nuclease (or another similar enzyme) degrades all single-stranded DNA, leaving only the duplex regions. NaPB, Sodium phosphate buffer.

Two assays suitable for use with tracers are nuclease digestion and hydroxyapatite chromatography. In nuclease assays, a single-strand-specific endonuclease such as S1 nuclease (Britten *et al.*, 1976) is used to digest the single-stranded DNA which remains in aliquots of the reassociation mixture sampled after various periods of incubation. The fraction resistant to digestion is taken as the fraction reassociated in each aliquot. Hydroxyapatite assays are based on the ability of hydroxyapatite to bind double-stranded DNA more tightly than single strands. Under certain conditions, such as 0.12 *M* sodium phosphate buffer at 60°C, single strands will pass through a column of hydroxyapatite while double strands are retained. Double-stranded material can then be removed by increasing the phosphate concentration, or eluted after being converted to single strands by raising the column temperature.

Both optical and nuclease assays measure the fraction of *nucleotides* which are paired. In contrast, hydroxyapatite measures the fraction of DNA *fragments* which have formed a duplex over any portion of their length, since the single-stranded regions of such fragments will be retained on the column by virtue of their attachment to the tightly bound duplexes. The minimum duplex length required for binding is usually considered to be about 50 nucleotide pairs (Wilson and Thomas, 1973).

Since reassociation experiments are normally conducted with randomly sheared DNA fragments, duplexes formed early in the reaction are seldom double-stranded over their entire length. As reassociation proceeds further, secondary reactions occur between the single-stranded "tails" of primary duplexes, so that fully reassociated DNA includes a high percentage of base pairs and may contain multimolecular structures involving several to many fragments. Secondary reactions are slower than primary reactions between free single strands, and complex kinetics are therefore observed when reassociation is monitored as the fraction of nucleotides paired (Smith *et al.*, 1975; Rau and Klotz, 1978). When assayed by hydroxyapatite chromatography, however, reassociation of simple DNA closely approximates second-order kinetics, as expected for a bimolecular reaction. This observation can be understood in terms of the fact that hydroxyapatite binding assays score fragments as reassociated as soon as the first duplex region is formed, thus providing a measure of primary or "first-collision" duplex formation which is relatively insensitive to slower reactions involving single-stranded tails.

Second-order reassociation kinetics may be described (Britten *et al.*, 1974) by the equation $C/C_0 = (1 + KC_0t)^{-1}$, where C_0 is the initial DNA concentration in moles of nucleotides per liter, C is the concentration of free single strands after t seconds of reassociation, and K is the rate constant of the reaction. It may be seen that, when $C/C_0 = \frac{1}{2}$, $K = 1/C_0t_{1/2}$. Reassociation data are frequently presented as C_0t curves (Britten and Kohne, 1968; C_0t is

usually verbalized as the acronym "Cot") such as that illustrated in Fig. 2. Plotting the results as a function of C_0t facilitates handling data obtained at several different DNA concentrations. The log scale is convenient, since even simple reactions cover more than 2 decades of C_0t and complex eukaryotic DNAs may reassociate over 5–10 decades.

Under a given set of reassociation conditions the rate constant for nonrepeated DNA is inversely proportional to the genome size (Britten and Kohne, 1968). This relationship can best be understood in terms of the concentration of each individual complementary sequence in the reaction. As the genome size increases, the concentration of any individual sequence in a given amount of total DNA decreases, decreasing the frequency of collisions between complementary sequence elements and hence the rate of reassociation.

Virtually all organisms contain DNA sequences whose reassociation kinetics suggest that they are represented in only one copy per haploid set of chromosomes. These are commonly called *single-copy, unique,* or *nonrepetitive sequences.* In addition, eukaryotic genomes typically contain substantial amounts of DNA which reassociates much more rapidly than expected for single-copy DNA and thus appears to contain families of related sequences present in multiple copies (Britten and Kohne, 1968). In most

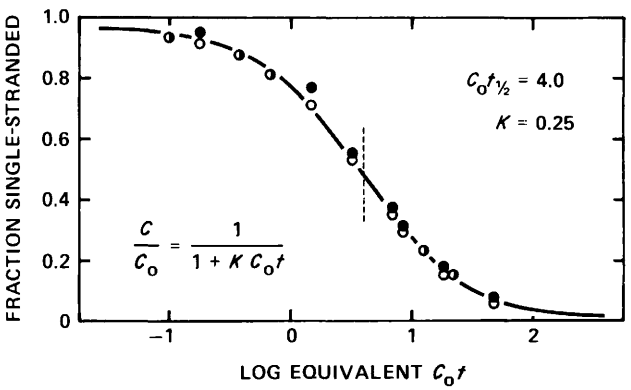

Fig. 2 Reassociation kinetics of two samples of *E. coli* DNA fragments as assayed by hydroxyapatite chromatography. The fraction remaining single-stranded is plotted as a function of C_0t, the product of DNA concentration in moles of nucleotides per liter and the incubation time in seconds. Even a simple DNA, lacking repeated sequences, reassociates over slightly more than two decades of C_0t, accounting for the fact that log scales are normally used in presenting such data. For a single component (i.e., a single-frequency class of sequences), reassociation assayed by hydroxyapatite closely approximates an ideal second-order reaction. The curve shown represents a component reassociating as predicted by the second-order rate equation at the lower left, with a rate constant of 0.25. By manipulating the rate equation, it may be shown that, at time $t_{1/2}$, when $C/C_0 = \frac{1}{2}$, $K = 1/C_0t_{1/2}$. More slowly reassociating DNA will thus have a lower rate constant and a higher $C_0t_{1/2}$ value.

instances such *repetitive sequences* fall into several classes on the basis of their relative reassociation rates, with apparent repetition frequencies sometimes reaching hundreds of thousands of copies. In the discussion that follows, we will consider various classes of sequences which can be distinguished by reassociation kinetic analysis and some aspects of their organization with respect to one another.

C. Single-Copy DNA

The most slowly reassociating sequences in a genome are usually assumed to be single-copy sequences, although proof of this assertion requires that the presumed single-copy component be shown to reassociate with the kinetics expected from the haploid genome size of the species. To do so, it is necessary to compare the observed reassociation rate with that for a standard DNA from an organism such as *Escherichia coli*. The ratio of reassociation rates should be that predicted from the ratio of genome sizes determined chemically. If so, the slowly reassociating sequences are present only once per haploid complement of DNA. Alternatively, if the reassociation of these sequences is faster than expected, it follows that they must actually be present more frequently than one copy per haploid genome.

A similar procedure can be used to derive an estimate of genome size for a species whose DNA content is unknown. It has even been claimed (Pearson *et al.*, 1978) that kinetic estimates are to be preferred over those based on direct chemical measurements. However, the use of kinetic data for such a purpose requires caution. The accuracy with which rate constants may be determined from total DNA may be severely limited when the single-copy fraction is small and poorly resolved, as it is in many plant genomes. This difficulty can sometimes be overcome by isolating a fraction enriched in single-copy DNA so that its reassociation may be more accurately measured, although contamination with slow repeat sequences may still be a concern. An even more serious problem derives from the fact that contaminants in plant DNA preparations can sometimes accelerate reassociation by severalfold (Merlo and Kemp, 1976; Murray and Thompson, 1977), leading to an underestimate of the true genome size. Kinetic estimates which fall below those based on chemical determinations should therefore be considered suspect, unless it can be established that no acceleration has occurred. This is done by including a small amount of labeled bacterial DNA as an *internal* standard in the plant DNA solutions and comparing its reassociation with that of the same bacterial DNA incubated alone. Internal standards are also useful as a means of correcting for viscosity effects, which tend to reduce reassociation rates in concentrated solutions of DNA (Britten and Kohne, 1968; Murray *et al.*, 1978a).

The majority of structural genes are composed of single-copy sequences in

all organisms for which data are available, and thus this fraction of the genome is considered to be of greatest importance in coding for proteins. Although some genes are known to be repetitive (see Section I,G,1), most polysomal mRNA and polyadenylated RNA sequences in several animal systems reassociate with single-copy DNA sequences (e.g., Goldberg *et al.*, 1973; Levy and McCarthy, 1975). Similar observations have recently been made for polyadenylated mRNA from several plant tissues (Goldberg *et al.*, 1978; Thompson *et al.*, 1979).

Although most structural gene sequences are unique, unique sequences are not always structural genes. Table I illustrates the range of single-copy DNA content in plants for which accurate measurements have been made. The amount of unique DNA does not correlate with biological complexity any more than total DNA content does. It is difficult to see why broad beans (*Vicia*) should need nearly twice as much structural gene DNA as peas (*Pisum*) and four times more than mung beans (*Vigna*), all of which belong to the family Leguminosae and are about equally complex in terms of structure and development. Nor is it easy to imagine that the number of coding sequences required to specify the proteins of different cereal grains should vary by as much as 100% even after allowing for ploidy differences. Similar discrepancies are obvious among animal genomes. For example, snails (*Limulus*) and toads (*Xenopus*) have almost as much single-copy DNA as mammals and about 20 times more than some insects (Davidson *et al.*, 1975a).

There is also direct experimental evidence suggesting that much of the single-copy DNA is not required to code for mRNA. When mRNA is hybridized to unique DNA under conditions designed to saturate all complementary portions of the DNA, only a small fraction (typically 1–10%) is found to be transcribed into mRNA (e.g., Kleiman *et al.*, 1977; Goldberg *et al.*, 1978). Nuclear RNA typically contains transcripts of 4–10 times as much DNA, but most of them never leave the nucleus and therefore do not code for proteins. For example, in gastrula-stage sea urchin embryos, about 28% of the unique DNA is transcribed into nuclear RNA, but only 2.7% codes for mRNA (Galau *et al.*, 1974; Hough *et al.*, 1975). Britten and Davidson (1976) have pointed out that, if the fraction of the total transcript used for mRNA remains constant, only about 10% of the unique sequences will be needed to produce mRNA even on the assumption that the entire genome is transcribed into nuclear RNA during the life cycle.

Several observations also suggest that most single-copy sequences are evolving rapidly and are therefore at least relatively free of the constraints imposed on coding sequences by natural selection (see Britten and Davidson, 1976). Direct tests of the hypothesis that most single-copy DNA evolves more rapidly than sequences which code for proteins have been made in two systems (Rosbash *et al.*, 1975; Angerer *et al.*, 1976). In both rodents and sea

TABLE I

Total and Single-Copy DNA Contents of Higher Plants

Species	Haploid DNA[a]			Single-copy DNA[b]		
	Ploidy	NTP × 10^{-9}	Ref.[c]	Percentage	Ref.[c]	NTP × 10^{-9}
Vigna radiata (mung bean)	2	0.5	1	70	1	0.4
Gossypium hirsuitum (cotton)	4	0.8	2	68	2	0.5
Glycine max (soybean)						
Embryo DNA	4	1.1	10	60	3	0.7
Leaf DNA				39	4	0.4
Petroselinum sativum (parsley)	2	1.9	5	30	5	0.6
Nicotiana tabacum (tobacco)	4	2.0	11	45	6	0.9
Avena sativa (oat)	6	4.3	11	25	7	1.1
Pisum sativum (pea)	2	4.8	11	30	8	1.4
Hordeum vulgare (barley)	2	5.4	11	30	7	1.6
Triticum aestivum (wheat)	6	5.6	11	25	7	1.4
Secale cereale (rye)	2	8.6	11	25	7	2.2
Vicia faba (broad bean)	2	13.0	11	20	9	2.6

[a] Where possible, values for DNA content have been taken from cytophotometric measurements. The genome sizes for mung bean and cotton were estimated from the reassociation kinetics of single-copy sequences. All values are listed in nucleotide pairs (NTP).

[b] Calculated from the indicated values for haploid genome size and percentage of single-copy sequences rounded to the nearest 0.1 × 10^9 NTP.

[c] References: (1) Murray *et al.*, 1979; M. D. Bennett, personal communication; (2) Walbot and Dure, 1976; (3) Gurley *et al.*, 1979; (4) Goldberg, 1978; (5) Kiper and Herzfeld, 1978; (6) Zimmerman and Goldberg, 1977; (7) Flavell *et al.*, 1977; (8) Murray and Thompson, 1978a; (9) W. C. Taylor and A. J. Bendich, personal communication; (10) Sparrow and Miksche, 1961; (11) Bennett and Smith, 1976.

urchins the sequences represented in mRNA show less divergence between related species than total single-copy sequences do, which strongly suggests that the former are conserved and the latter are free to vary. It is possible to calculate that a majority of the single-copy DNA is composed of the more variable sequences and thus does not appear to have the properties expected for functional genes.

It appears, therefore, that most of the single-copy fraction is free to vary quite rapidly in evolution, both in base sequence and total amount. However, Fig. 3 shows that a remarkably consistent relationship can be demonstrated between the amount of single-copy DNA and the genome size in different plant and animal species. The apparent consistency of this relationship is surprising, as is the clear difference between plants and animals which becomes apparent at genome sizes above about 10^9 nucleotide pairs. As a working hypothesis, it is possible to suggest that single-copy sequences accumulate in larger plant and animal genomes by gradual evolutionary divergence of repeated sequences. On this hypothesis, the members of a repeated family would become progressively more and more different from one another until the differences were so great as to prevent their cross-reassociation under the usual *in vitro* conditions. The different slopes of the curves for plants and animals might then be attributed to a higher average rate of repeated-sequence amplification in plant genome evolution. Given similar rates of sequence divergence in plants and animals, the more fre-

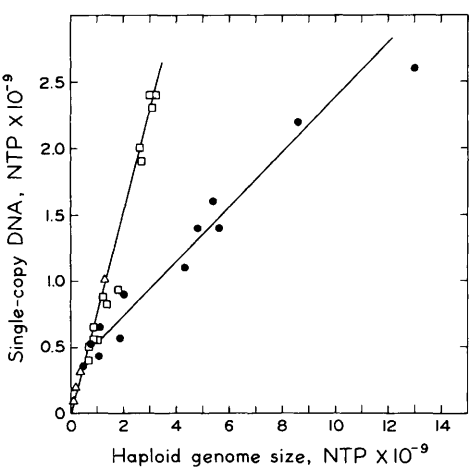

Fig. 3 Single-copy DNA content as a function of total genome size in plants (solid symbols) and animals (open symbols). Data for plants are from Table I; those for animal genomes were taken from Davidson *et al.* (1975a), Crain *et al.* (1976), Efstratiadis *et al.* (1976), Sachs and Clever (1972), Schmid and Deininger (1975), and Arthur and Straus (1978). Values for single-copy DNA are based on measurements (such as S1 nuclease resistance or optical hyperchromicity) not influenced by repetitive sequence interspersion. Open triangles indicate animal genomes with long-period interspersion patterns (see Section I,E).

quent addition of repeats would lead to a higher steady-state percentage of repetitive DNA.

The gradual divergence hypothesis predicts a more-or-less continuous distribution of divergence among groups of repetitive sequences originally amplified at different times in the evolutionary history of a species, leading to a similar distribution in the precision of base pairing in reassociated repetitive duplexes. Reassociated repetitive sequences in most eukaryotes melt over a wide range of temperatures, indicating that broad distributions of pairing precision exist; this phenomenon is discussed in more detail below in connection with repetitive-sequence organization. However, a related prediction is that the fraction of the genome reassociating with single-copy kinetics should be a function of the stringency of the reaction conditions. As the stringency is increased, e.g., by raising the reassociation temperature, the most divergent repetitive sequences will no longer be able to cross-react with one another. However, they would still be able to form duplexes with their exact complements from the opposite DNA strand and would therefore reassociate at the rate expected for single-copy sequences. The reverse effect should also be observed as the temperature is lowered, provided that the distribution of sequence divergence extends to sufficiently low levels of homology.

In accordance with this prediction, Bendich and McCarthy (1970) observed that the fraction of wheat DNA reassociating with single-copy kinetics varied between 0 and 50%, depending on the stringency of reaction conditions. Very similar results have recently been obtained by Flavell and Smith (1976) for wheat and by Smith and Flavell (1977) for rye DNA. Bendich and Anderson (1977) showed that the amount of DNA in the most rapidly renaturing major repetitive components of DNA from four unrelated plant species exhibited a similar variation with reassociation criterion. Thus a large fraction of the sequences in these genomes can be recognized as repetitive only under relatively permissive reaction conditions. In mouse DNA, Ivanov and Markov (1978) have recently shown that some sequences which appear to be unique under standard conditions can reassociate with the repetitive fraction when the temperature is lowered. While the evidence is not yet extensive, there is a general indication of the absence of a sharp distinction between single-copy and repeated DNA consistent with the hypothesis that single-copy sequences are continuously generated by divergence within repetitive families during the course of evolution.

D. Repetitive DNA

1. Repetition Frequencies

Given an accurate measurement of the reassociation rate for a group of repetitive sequences, it is possible to calculate an apparent repetition fre-

quency by comparing this rate with the rate for single-copy sequences in the same genome. When the observed reassociation kinetics are complex, indicating the presence of more than one repetition frequency class, it is usual to analyze the data with a computer program (e.g., Pearson *et al.*, 1977) designed to fit a number of ideal second-order components to the data. Table II summarizes the results of such fitting procedures for DNA from a number of higher-plant species. It is immediately apparent that the range of repetition frequencies encountered is broad, with values ranging from between 20 and 300,000 copies per haploid genome. In most cases, however, components accounting for a significant fraction of the total DNA seem to be present in a few thousand copies or less.

Table II also lists estimates of the *kinetic complexity* for each repetitive component. Kinetic complexity is defined (Britten *et al.*, 1974) as the total length of DNA contained in *different* nucleotide sequences as measured by reassociation kinetics. It thus provides a measure of information content not affected by repetition frequency. For a given component, kinetic complexity may be calculated by comparing its reassociation rate constant to that of a standard with known complexity such as *E. coli* DNA. This procedure requires that rate constants observed in unfractionated whole DNA be divided by the fraction of the genome in a given component in order to obtain the rate which would have been observed for that component in the pure state ($K_{\text{pure}} = K_{\text{whole}}/\text{fraction}$). Thus a component including 50% of the genome and reassociating in total DNA with a rate constant of 1.0 would have a K_{pure} of 2.0 (= 1.0/0.5). If *E. coli* DNA reassociated under the same conditions had a rate constant of 0.2, this hypothetical component would have a kinetic complexity about 10 times lower than that of *E. coli* (2.0/0.2 = 10), or 4.5×10^5 nucleotide pairs.

It is clear from Table II that some repetitive sequences in plant genomes have relatively high kinetic complexities, and thus potential information contents. In several cases the kinetic complexity reaches several million nucleotide pairs, a value roughly equivalent to the size of a bacterial genome. At the other extreme, components with much lower complexities also exist in many plant genomes. For example, Goldberg (1978) reported about 3% of soybean DNA to be composed of sequences repeated about 300,000 times, with a kinetic complexity of only about 200 nucleotide pairs. This small component was not resolved when total DNA reassociation was analyzed but could be detected in a fraction enriched for highly repetitive sequences. Additional evidence for high-frequency, low-complexity components in other plant genomes not listed in Table I comes from studies on density satellites, which will be reviewed below.

While the data in Table II provide a useful indication of the range of repetition and complexity found among higher-plant DNAs, it is important to recognize that the values listed are at best crude approximations. For exam-

TABLE II

Component Analysis of Repetitive DNA Reassociation Kinetics[a]

Species	Reference[b]	Component	Fraction	Copies	Kinetic complexity (NTP)
Vigna radiata (mung bean)	1	Very fast[c]	0.02		
		Fast	0.05	10,000	2.8×10^3
		Slow	0.28	175	9.0×10^5
Gossypium hirsuitum (cotton)	2	Very fast	0.08		
		Repeats	0.27	130	1.6×10^6
Glycine max (soybean)					
Embryo DNA	3	Very fast	0.05		
		Fast	0.28	3,850	9.4×10^4
		Slow	0.24	153	2.0×10^6
Leaf DNA	4	Very fast	0.07		
		Fast	0.28	2,784	2.2×10^5
		Slow	0.38	19	1.8×10^7
Petroselinum sativum (parsley)	5	Very fast	0.05		
		Highly repetitive	0.13	136,000	1.8×10^3
		Fast	0.48	3,000	3.0×10^5
		Slow	0.16	42	7.4×10^6
Nicotiana tabacum (tobacco)	6	Very fast	0.02		
		Fast	0.06	12,400	4.6×10^3
		Intermediate	0.59	252	2.5×10^6
Pisum sativum (pea)	7	Very fast	0.04		
		Fast	0.46	10,000	2.0×10^5
		Slow	0.33	300	5.0×10^6

[a] All data are derived from experiments in which the reassociation of short fragments of total cellular DNA was monitored by hydroxyapatite binding. Fragment lengths varied between 250 and 470 nucleotides in different reports.

[b] References: (1) Murray *et al.*, 1979; (2) Walbot and Dure, 1976; (3) Gurley *et al.*, 1979; (4) Goldberg, 1978; (5) Kiper and Herzfeld, 1978; (6) Zimmerman and Goldberg, 1977; (7) Murray *et al.*, 1978a.

[c] The very fast components for all species contain foldback (inverted repeat) DNA but in some cases may also contain a fraction of very highly repeated sequences which reassociate before the earliest data points obtained.

ple, the estimates of repetition frequency do not take into account the effects of mismatched base pairs on reassociation rates. Repetitive sequence families may exhibit considerable sequence divergence so that the members no longer reassociate with one another to produce precisely paired duplexes; in many cases the average amount of mispairing observed in duplexes formed under standard conditions is on the order of 10%. This amount of mismatch has been shown to reduce the reassociation rate by a factor of about 2 (Bonner *et al.*, 1973; Marsh and McCarthy, 1974), and even larger effects would be anticipated for duplexes with greater than average mismatch.

Whether or not plant DNA sequences are actually present in a small number of discrete frequency classes, as Table II suggests, is also questionable. C_0t curves for many plant DNAs are nearly linear throughout the range in which repetitive sequences reassociate, as illustrated for pea DNA in Fig. 4. Although computer fitting procedures can usually be used to model these curves with two or three ideal components, the mere construction of such models does not constitute evidence that the theoretical components actually exist as discrete entities. Unless discrete components can actually be

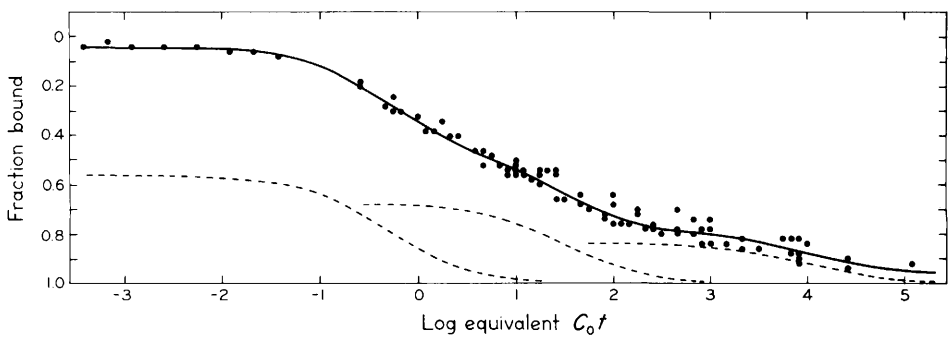

Fig. 4 Reassociation kinetics of short pea DNA fragments. DNA fragments sheared to a modal single-strand length of 300 nucleotides were dissolved in 0.05, 0.12, or 0.40 *M* sodium phosphate buffer (pH 6.8) or in 1.0 *M* sodium chloride–10 m*M* sodium phosphate buffer at concentrations ranging between 250 and 5000 μg/ml. All reassociation buffers contained 0.1 m*M* EDTA. Aliquots of each solution were sealed in capillary tubes, heat-denatured, and reassociated at 55°, 60°, 66°, or 70°C to provide a constant reassociation criterion at all four salt concentrations. Reactions were assayed by hydroxyapatite fractionation at 60°C in 0.12 *M* sodium phosphate. All C_0t values have been corrected to equivalent C_0t values in 0.12 *M* phosphate buffer at 60°C (see text). ^{3}H-labeled pea DNA fragments were used at both extremes of the curve, where the small amount of DNA in either the single-stranded or double-stranded fraction makes accurate absorbance measurements difficult. The solid line is a least squares fit to the data using three theoretical second-order components (shown by dashed lines) plus a 4% "very fast" fraction reacting prior to a C_0t of 0.01. The root mean square error of the indicated least squares fit was 0.031. From Murray *et al.* (1978a).

isolated in pure form, such computer models must be regarded merely as conveniently simplified mathematical descriptions of the data. Relatively few attempts have been made to isolate discrete components from plant DNA, and they have usually failed to produce convincing evidence. A DNA fraction whose reassociation approximates that expected for a single ideal component has been obtained from tobacco DNA (Zimmerman and Goldberg, 1977). However, further analysis of an apparently simple repetitive component in mung bean DNA has shown it to be composed of at least two quite different frequency classes (Preisler and Thompson, 1978; Murray *et al.*, 1979). Similar experiments with repetitive fractions from the pea genome have indicated even more kinetic heterogeneity (Fig. 5 and Murray *et al.*, 1978a). In this case, the results are best interpreted by assuming an essentially continuous distribution of repetition frequencies instead of the simple two-class model derived from computer fits to total DNA data. Working with individual repetitive sequences cloned in *E. coli* by recombinant DNA techniques, Klein *et al.* (1978) were recently able to show that the major repetitive component of the sea urchin genome was also composed of a large number of different repetition frequency classes.

2. Satellite DNA

The term "satellite DNA" refers to nuclear DNA sequences which can be separated from the main component DNA of a genome by physical methods, most commonly by equilibrium buoyant density centrifugation in gradients of cesium salts (Skinner, 1977). Satellites have been found in the genomes of a wide variety of plants and animals. Among higher plants, satellites detectable by neutral cesium chloride centrifugation were found by Ingle *et al.* (1973) in 27 of the 59 dicot species they examined. The fraction of total DNA in an individual satellite varied from 3 to 28%, and two satellite components found in cucumber DNA represented about 44% of the nuclear DNA. No satellites were observed in any of the 11 monocots species examined in this study, nor in the 15 monocots analyzed subsequently (Ingle *et al.*, 1975). However, in the latter survey Ingle *et al.* demonstrated an inverse correlation between genome size and the presence of detectable satellite DNA in neutral cesium chloride. Since the monocot species surveyed all had relatively large genomes, the apparent absence of satellite sequences in these species probably reflects the dificulty of detecting small components in a large amount of total DNA rather than any fundamental difference between monocot and dicot plants. An A+T-rich satellite has subsequently been found in at least one monocot, the orchid *Cymbidium* (Capesius *et al.*, 1975).

In addition to satellites observable in neutral cesium chloride gradients ("patent" satellites; Skinner, 1977), other satellite DNAs may be revealed by centrifugation in cesium sulfate gradients containing heavy-metal ions (such as silver or mercury) or antibiotics (such as actinomycin D) which bind

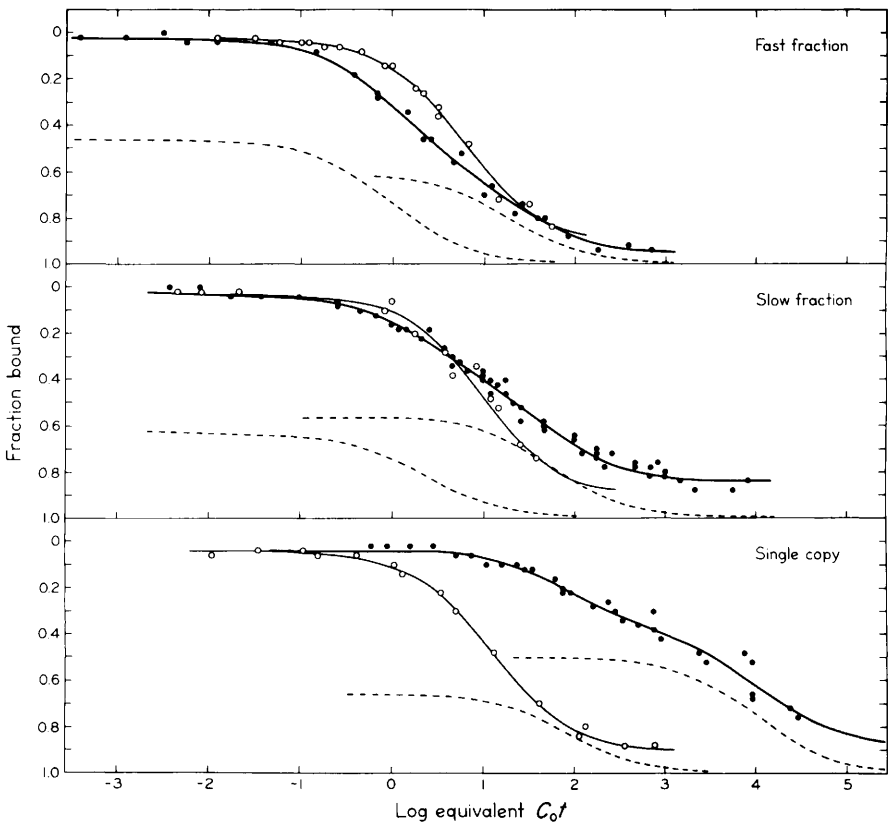

Fig. 5 Reassociation of isolated kinetic fractions of the pea genome. Three-hundred-nucleotide pea DNA fragments were separated into three fractions by sequential reassociation and fractionation on hydroxyapatite. The reassociation histories (Britten *et al.*, 1974) and fractions of total DNA were fast repeat: C_0t 200 bound, C_0t 2 bound, and C_0t 0.01 unbound (24%); slow repeat: C_0t 200 bound and C_0t 2 unbound (42%); single-copy: C_0t 200 unbound and C_0t 1000 unbound (13%). ^{14}C-labeled tracers were prepared from each fraction with *E. coli* DNA polymerase I (Murray *et al.*, 1977, 1978a). Each tracer was mixed with a 2000-fold excess of unlabeled fragments of total DNA, and aliquots were then reassociated and fractionated as in Fig. 4. In each panel, the heavy solid line represents the best two-component least squares fit to the pea tracer data. The dashed lines depict the theoretical second-order components of these hypothetical solutions. The light solid lines show the reassociation kinetics of ^{3}H-labeled *E. coli* internal standards in each reaction. *Escherichia coli* rate constants were 0.194, 0.114, and 0.091 in the fast, slow, and single-copy reactions, respectively, reflecting the retardation of reassociation by the higher viscosities of the more concentrated DNA solutions used to achieve high C_0t values. From Murray *et al.* (1978a).

differentially to different DNA sequences. Such cryptic satellites can frequently be found in genomes lacking conventional satellite DNA. Patent satellites may sometimes be resolved into two or more components by similar techniques.

Although satellite DNAs are frequently analyzed in preparations of total cellular DNA from plant tissue, it is unlikely that organelle DNA contributes significantly to most of the satellite fractions normally studied. Chloroplast DNA is generally not distinguishable from main-band DNA in cesium chloride gradients (Wells and Ingle, 1970; Pascoe and Ingle,1978), and mitochondrial DNA, although banding at a density similar to that of many plant satellites, has been reported to contribute less than 1% of the total DNA mass (Ingle *et al.*, 1973). Ribosomal cistrons might be expected to form a dense satellite if present in sufficient quantity, and at least some of the DNA in satellite fractions from several plants has been identified as rDNA. However, Ingle *et al.* (1975) have shown that, although rDNA and satellite DNA densities are often similar, it is usually possible to distinguish between them, indicating that rDNA generally comprises a small fraction of the total satellite. In the case of the muskmelon (*Cucumis melo*) these authors showed that the satellite DNA sequences were not covalently linked to ribosomal cistrons. The greatest enrichment for ribosomal cistrons yet demonstrated in a higher-plant satellite DNA is found in the tomato satellite studied by Chilton (1975). About half of the DNA in this satellite, or 3% of the genome, can be identified as ribosomal cistrons.

Satellites are usually observed in experiments with relatively long DNA fragments (10,000 nucleotide pairs is fairly typical), and thus DNA molecules banding in the satellite region of the gradient must contain long stretches of sequence(s) whose base composition or other physical properties differ from the main-band DNA in some unique way. On a priori grounds, it seems most likely that long strands with such uniformly unique properties would contain many tandem repeats of a relatively simple, distinctive sequence. Most satellites so far examined contain tandem repeats, and in animals virtually all are very simple sequences with kinetic complexities of a few hundred nucleotide pairs or less (Skinner, 1977). Sequencing studies on animal satellites have revealed several cases in which the basic repeating unit is only a few bases in length. In the hermit crab, for example, satellite I is composed almost exclusively of the sequence TAGG and its complement, while satellite II contains repeats of the sequence AGTGCAG$(CTC)_n$, where $n = 3–12$ (Chambers *et al.*, 1978). Perhaps the simplest satellite known is that of the land crab, which is more than 90% poly[d(A-T)]. Very simple repeating units may also be present in more complex satellites. For example, in mouse satellite DNA the sequence GA_nTGA (where n is about 5 or less, on the average), and variants produced from it by a single base change, account for about 50% of the total. Higher-order periodicities of about 250 nucleotide

pairs have also been identified by restriction enzyme analysis (see Flashman and Levings, this volume, Chapter 2) of the mouse satellite, leading to the conclusion that this DNA might have originated from a simple oligonucleotide repeat which became extensively modified during evolution. More extensive discussion of animal satellite DNAs can be found in Skinner (1977).

Sequencing techniques have not yet been applied to plant DNA satellites. However, studies on reassociation kinetics have shown that much more complex sequences are commonly found in plant satellites than in those of animal genomes. It is also interesting to note that all the plant satellite DNAs so far examined in detail contain two or more components which, although banding at the same density in neutral cesium chloride, are distinguishable on the basis of their thermal denaturation profile and/or reassociation kinetics.

Perhaps the best known satellite DNA from a higher plant is that of the muskmelon (*C. melo*) studied by Bendich and Anderson (1974), Sinclair *et al.* (1975), and Bendich and Taylor (1977). The isolated satellite forms a homogeneous band at 1.707 g/ml in neutral cesium chloride. However, thermal denaturation profiles clearly show at least two components. About two thirds of the purified native satellite DNA melts with a T_m of about 86°C is SSC (0.15 *M* sodium chloride, 0.015 *M* sodium citrate, $Na^+ = 0.195$ *M*), while the remaining one-third melts with a T_m about 8°C higher. Applying conventional formulas for the relationships between T_m, G+C content, and buoyant density, Sinclair *et al.* calculated that the low- and high-melting fractions of the satellite should have densities in neutral cesium chloride of 1.704 and 1.725, respectively. In order to explain the fact that these two components band together at 1.707 g/ml, it is therefore necessary to suppose that they are covalently linked to one another, or that the high-melting fraction has an extremely anomalous density (Bendich and Anderson, 1974). Although the latter hypothesis might seem less likely on a priori grounds, both Bendich and Anderson (1974) and Sinclair *et al.* (1975) showed that the two components could in fact be separated by centrifugation in cesium sulfate gradients containing heavy-metal ions, or in cesium chloride after denaturation and partial renaturation. Bendich and Taylor (1977) exploited the large difference in T_m between the two components to demonstrate the absence of linkage by electron microscopy. At an appropriate temperature between the two thermal transitions, essentially all the molecules (40,000 nucleotide pairs in length) were either completely denatured or completely native. Partially denatured, partially native molecules should have been seen if the two fractions were covalently linked, and thus this experiment conclusively eliminated the linkage hypothesis.

Muskmelon satellite also contains two principal components by renaturation analysis. The high-melting fraction renatures at a rate corresponding to a

kinetic complexity of 600 nucleotide pairs and an apparent repetition frequency of 1.8×10^5 copies per haploid genome, while the low-melting fraction has a kinetic complexity of 1.8×10^6 nucleotide pairs and is represented in about 70 copies (Sinclair *et al.*, 1975; Bendich and Taylor, 1977). The complexity of the latter component is about 40% of that of the *E. coli* genome, higher than that for any other known satellite sequence. Both components renature almost perfectly, in contrast to many other plant and animal satellite DNAs. The thermal stability of the high-melting, low-complexity fraction was reduced by only about 1°C, while that of the low-melting, high-complexity component was essentially unchanged (Bendich and Anderson, 1974; Sinclair *et al.*, 1975). In this case, therefore, estimates of sequence complexity based on reassociation kinetics are not complicated by the effects of base pair mismatching within reassociated duplexes. The extremely high degree of base pairing achieved also appears to rule out a pattern of sequence organization in which only very short segments of the molecules actually form duplexes. Thus muskmelon satellite DNAs would not be expected to exhibit the anomalous reassociation behavior observed for mouse satellite DNA (Chilton, 1973; Hutton and Wetmur, 1973).

Most plant DNA satellites characterized by reassociation also contain two or more kinetic components, and all except the flax and wheat satellites exhibit complex thermal denaturation profiles (Table III). Only in the case of the tomato satellite, where the more complex component appears to code for rRNA (Chilton, 1975), has it been possible to assign a definite function to any of the satellite sequences. The nearly universal occurrence of complex sequences in plant DNA satellite fractions is in marked contrast to the general picture for animal satellites, but the significance of this fact remains to be determined. Since the complex components account for 40–50% of the satellite preparations, it is difficult to attribute them to simple contamination with residual main-band DNA fragments. In most cases it is possible to imagine that the different classes of sequences might be covalently linked to one another in the genome so that they band together at the same density. As noted above, however, this possibility has been ruled out in the case of the muskmelon satellite.

E. Interspersion of Repetitive and Single-Copy Sequences

In contrast to satellite DNA, most repetitive sequences are interspersed to varying degrees with single-copy sequence elements. As early as 1965, Bolton *et al.* observed that long fragments of eukaryotic DNA could form large multimolecular aggregates or networks after partial reassociation and concluded that the repetitive sequences responsible for base pairing under their conditions must be scattered throughout the genome. Experiments with mouse DNA fragments of several different lengths suggested that repetitive