J Mol Evol (1986) 24:12–27

Journal of
Molecular Evolution

Polite DNA: Functional Density and Functional Compatibility in Genomes*

Emile Zuckerkandl

Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, California 94306, USA

Summary. Certain as yet poorly defined functions of DNA appear to involve collectively domain-sized sequences. It is proposed that most sequence segments within a domain may be either functionally superfluous or instrumental, depending on how many related sequences are present in the domain. When redundant and functionally dispensable, such DNA segments presumably still have to conform to compositional or sequence-motif patterns that characterize the domain. In its relations with neighboring sequences, such DNA is required to be "polite." Polite DNA is DNA that, without being crucially involved in function, is subject to constraints of conformity and, through its base composition, respects a function for which it is not required. This concept is developed by contrasting the distribution of specific and general functions over DNA with this distribution as found in proteins and by distinguishing functional compatibility from pivotal functionality. The sequence constraints to which heterochromatin as well as, apparently, long interspersed repetitive sequences are known to be subject seem to imply that DNA, even when it does not carry out a pivotal function, is indeed, at the very least, required to be polite.

Key words: Polite DNA — Noncoding sequence functions — Sequence motifs — Junk DNA — Natural selection

Introduction

The structure of contemporary genomes can be interpreted either primarily as an evolutionary build-up of functional features or primarily as accumulated effects generated by spontaneous and nonselected processes. How much function and how much nonfunctional effects have contributed to the structure of contemporary genomes is one of the most hotly debated questions in molecular evolution today.

The tendency has been to consider that eukaryote DNA is divided between sites involved in specific functions (coding sequences and specific cis-acting noncoding sequences endowed with regulatory roles) and a large majority of functionless sites. Nowadays this extreme view is less widely held. Even the proponents of the "neutral theory" are not lagging behind others in accepting, indeed emphasizing, the evidence in favor of evolutionary constraints imposed by negative selection, constraints that are seen to affect increasingly numerous and diverse parts of the genome (Kimura 1983). There was a time when evolutionary modifications in certain fast changing parts of proteins, notably the fibrinopeptides, were considered to reflect the true neutral mutation rate. Driven from this position, first by Barnard et al. (1972), the neutral mutation rate then held on to the third codon positions, in spite of the fact that there had been early warnings to the effect that this was an uncomfortable location for it to lodge in (Zuckerkandl 1965; Zuckerkandl and Pauling 1965). Having emigrated again, the neutral mutation rate landed on the pseudogenes, perhaps once again imprudently (Vanin et al. 1980). There is no telling when and where, if anywhere, the peregrinations through the genome of the true neutral mutation rate will end.

That functionality may be pervasive in vertebrate genomes is suggested by the recent observation that these genomes are divided into a certain number of compositional sectors, in each of which compositional characters at the third codon positions closely

Offprint requests to: Emile Zuckerkandl

* Presented at the FEBS Symposium on Genome Organization and Evolution, held in Crete, Greece, September 1–5,1986

match compositional characters in flanking sequences outside of the genes (Bernardi et al. 1985; Bernardi and Bernardi 1985, 1986, 1987). In each sector the constraints imposed upon third codon positions (Jukes 1978; Grantham et al. 1980) and noncoding sequences are, on the average, essentially the same. Moreover, during their evolution, eukaryote genomes, or significant sectors thereof, are seen to undergo directional changes in base composition. Once these changes have occurred, the new compositional patterns would appear to be conserved over longer evolutionary periods than would be expected if the compositions and the sequences to which they correspond were changed largely by random drift. Though random drift might still be frequent, there probably are other, superimposed processes.

These superimposed processes might, to be sure, be attributable to something other than functionality. Gene conversion (Zimmer et al. 1980) and processes leading to related effects subsumed under the term molecular drive (Dover 1982); or sectorially variable trends in introducing replicational errors; or mutagenic effects differing qualitatively in different parts of the genome; or regional variations in the efficiency of DNA repair mechanisms (Madhani et al. 1986) all might be invoked though mostly with limited plausibility, to counter a selectionist interpretation of the results of Bernardi and his associates. However, the correlation found by Bernardi and Bernardi (1986, 1987) in poikilothermic animals between GC content and ambient temperature considerably strengthens the case in favor of a functional connection for GC content.

In this paper I wish further to explore the functional connections of most of the noncoding sequences and the concept of their functionality itself.

Evidence for Functional Involvement of the Bulk of Noncoding Sequences

We can expect that we are dealing with underlying functions rather than with nonselected effects if two conditions can be shown to hold:

(1) that evolutionary random decay would have effaced the observed regularities over the evolutionary time elapsed since these regularities appeared, and

(2) that these regularities are so structured that they cannot easily be explained by nonselected processes that would restore them after they decayed.

Over significant sectors of noncoding DNA sequences regularities have been found that, I believe, tend to satisfy both these conditions.

Systematic regularities can have the form of defined sequences; or of sequence motifs; or of compositional features. Recurrent compositional features will automatically lead to recurrent sequence motifs, provided the latter are given enough degrees of freedom of variation. Sequence motifs are defined as short runs of nucleotides, a certain fraction of which is free to vary from repeat to repeat, while another fraction, not always in the same positions, remains constant (Zuckerkandl 1981).

One of the main pieces of evidence for systematic regularities is sequence features that are instrumental in the precise positioning of nucleosomes. It was originally thought that sequence specificity did not exist in the case of the histones and of the nucleosomes that they build up (Prunell and Kornberg 1978). More recently, however, a sequence-determined phasing of nucleosomes has been observed. This phasing may be due either to an organizing boundary (Kornberg 1981) or to reiterated sequence features, and it seems that, at least for certain genes, the latter is the case (Benezra et al. 1986). Similar data relate to some nonhistone chromosomal proteins (NHCPs). In certain regions of primate genomes a High Mobility Group (HMG) protein, α-protein, appears to bind to nucleosomes on the basis of a direct interaction with a reiterated sequence motif whereby the nucleosomes are precisely positioned along the DNA (Strauss and Varshavsky 1984).

The sequence characters of DNA that define the spacing of the nucleosomes recur every 190 or 200 bp in mammals. This has perhaps most convincingly been shown for the mouse β-globin gene region of DNA (Benezra et al. 1986). The findings are in accord with unpublished observations that Dr. Steve Burbeck made in 1982 at the Linus Pauling Institute (Palo Alto, CA) by studying sequence motif periodicities in DNA with the help of a Fourier transform method that he had worked out. Similar periodicities have been noted in "nude" eukaryotic DNA by Udvardy and Schedl (1983) in regard to the methylation of cytosine residues, and again in nude DNA by Keene and Elgin (1984), who found the periodicities in the form of preferred cleavage sites, using micrococcal nuclease or, with the same result, the intercalating 1,10-phenanthroline-cuprous complex. The latter authors have analyzed noncoding sequences in the vicinity of 18 *Drosophila* genes and of 2 mammalian genes and noted consistently similar cleavage patterns. While not evident in coding sequences, the patterns extend to the larger intervening sequences.

It had been predicted that intervening sequences would participate in the pattern of periodic repeats of sequence motifs (Zuckerkandl 1981), the idea being that such reiterated sequence motifs would function in the binding of certain proteins, thought to carry out certain structural and regulatory func-

tions. It had also been predicted that such periodic binding sites would not generally be found within the coding sequences, and it was proposed that a general function of intervening sequences was to break up the continuity of coding sequences so that they could not seriously interfere with the stability of higher order polynucleotide structures. Data now at hand are compatible with these views.

Certain regularities of sequence features of DNA are difficult to explain in terms of nonfunctional effects of self-generated sequence multiplication (Doolittle and Sapienza 1980; Orgel and Crick 1980) and sequence correction (Dover 1982) processes, or of nonselected DNA slippage leading to "cryptic simplicity" (Tautz et al. 1986). Interestingly, we may here bring in satellite DNA, which in the eyes of not a few is as good "junk" DNA (Ohno 1972) as junk DNA comes. The HS-α satellite DNA in rodents (Fry and Salser 1977) is an example of a highly repetitive simple sequence present in species whose common ancestor is distant enough for the contemporary sequences to be expected, if they freely accept mutations, to differ from each other more than they do. Such sequence invariance does suggest, though not demonstrate, functional involvement. Further evidence in support of some function of satellite DNA is provided by differences in frequency of sequence divergence at different nucleotide positions in the basic repeat unit of several satellites (Brutlag 1980). Sequence can be highly conserved at certain nucleotide sites and quite variable elsewhere.

The strongest evidence in support of the concept that satellite DNAs—at least some of them—must conform to constraints is provided by the finding that not only does α-satellite in cells from the African Green Monkey display sites for the binding of a specific nonhistone chromosomal protein, the α-protein already mentioned, but these sites are so spaced as to span the length of the core of a nucleosome on the one hand (145 bp) and that of linker DNA (27 bp) on the other (Strauss and Varshavsky 1984). One is led to infer that evolutionary constraints bear on the sequence of this satellite so as to bring about reiterated binding to DNA, with appropriate regular spacings, of the HMG-type protein and thereby to ensure a proper phasing of nucleosomes. If satellite DNA is junk, it is, at the very least, neatly packaged junk. The packaging can hardly be maintained over evolutionary time without the intervention of at least negative selection. It is of course possible and I would say credible, Miklos and Gill (1981) notwithstanding, that satellite DNA is neatly packaged because it has a specific function, such as to fix interphase chromosomes at certain points in the nuclear membrane (see Hutchison and Weintraub 1985).

Periodically recurring sequence motifs generated by spontaneous processes with no connection to natural selection would be supposed to accept neutral mutations indefinitely. The periodicities would be effaced at the rate of at least 1% fixed nucleotide substitutions per million years.[1] It is asking rather much of neutrality to maintain or restore sequence characters in such a way that they happen to coincide quite generally and to remain coincident with the period of nucleosomes. That functional neutrality is not the master of the field of the noncoding sequences is also suggested by the constraints spread over the approximately 60 kb of the anthropoid β-globin gene complex. This complex is evolutionarily rather stable in its overall organization and accepts substitutions at a rate of only about 0.2% per million years (Barrie et al. 1981; Jeffreys 1982). Neither the coding sequences alone, which represent about 8% of this complex, nor these sequences in combination with the small fraction of highly constrained cis-acting regulatory sequences can be held responsible for most of the slowdown relative to what is considered the neutral substitution rate.

When rather distantly related species are being compared, as is the case of the comparison between the α-satellites of kangaroo rat, guinea pig, pocket gopher, and antelope ground squirrel (Fry and Salser 1977), one would expect any "master" sequence from which "slave" sequences are generated or, if preexisting, are corrected, to be itself the object of mutational change. Master and slave sequences: I am using Callan's (1967) old terminology. It is unlikely that the conservation of master sequences could take place without the intervention of natural selection and the presence of a function. Admittedly, in selfish DNA (Doolittle and Sapienza 1980; Orgel and Crick 1980), sequences necessary for its maintenance would be expected to be conserved over long evolutionary periods. The structure of satellite DNA however does not encourage one in the thought that satellite DNA may be selfish. When selfish DNA does occur, what is conserved presumably is the sequence apparatus for selfish replication and transposition, not any further sequence features that spread through the genome with the help of this apparatus. Furthermore, if selfish DNA sees to it, as an expression of its selfishness, that it remains in a state proper for its own packaging, it ceases to be altogether self-centered in that it adapts itself to the genome that it inhabits.

---

[1] From the data of Miyata and Yasunaga (1981) and those of Li et al. (1981), discussed by Kimura (1983), and considering that base substitutions in pseudogenes may *not* represent quite the true neutral mutation rate, as suggested by the work of Bernardi and collaborators, we may consider that a figure for the true neutral mutation rate of 1% evolutionarily effective substitutions per million years probably is on the low side.

Table 1. Functional effects of DNA.* The variable relationships between function and the nucleotide substratum of function lead one to distinguish between (1) specific functional effects involving specific individual nucleotides, (2) specific functional effects obtained on the basis of low specificity of individual nucleotides and probably depending on compositional features reflected in recurring sequence motifs, and (3) general functional effects obtained on the basis of low or very low specificity of individual nucleotides

A. Specific functional effects based on sequence constraints of high to moderate specificity
–coding
–transcriptional regulation based on promoters, enhancers, silencers (Laimins et al. 1986), homoeoboxes, topoisomerase II cleavage sites (Udvardy et al. 1985), functions linked to hypersensitivity to endonucleases (Larsen and Weintraub 1982; Weisbrod 1982; Weintraub 1983), etc.
–regulation of processing and translation based on specific sequence features of RNA transcripts, for instance regulation of translation rate by choice of nucleotide at the "silent" codon position (Zuckerkandl 1965; Ikemura 1985)
–regulation of replication based on short highly or fairly specific DNA sequences
–various functions linked to the binding by DNA, RNA, or by proteins of small RNAs, as required e.g. for processing of RNA transcripts (Schaufele et al. 1986)
–functions depending on rather specific local structures of DNA or RNA such as, in DNA, hairpin loops, and with a lesser degree of sequence specificity, the formation of z-DNA, and perhaps sequence-directed curvature (Hagerman 1986; Koepsel and Kahn 1986). Certain functions linked to hypersensitivity to endonucleases
–aspects of transcription and replication depending on specific attachment sites of DNA domains to the nuclear matrix (Moreau et al. 1982; Goldberg et al. 1983; Mirkovitch et al. 1984; Cockerill and Garrard 1986)
B. Specific functional effects based on sequence constraints of moderate to low specificity
–mass binding of certain proteins (see text), such as histone H1, insofar as it provides the chromosomal and nuclear structural basis for different levels of regulation of gene expression as well as for DNA replication
–various functions with often as yet ill-defined structural bases, relating for example to interactions between chromosomes and the nuclear membrane (e.g., Hillicker and Appels 1980), to chromosome pairing, to recombination, to the functioning of centromeres and telomeres (Holmquist and Dancis 1979; Hutchison and Weintraub 1985), etc.
–regulation of cell size as a function of the c-value (Cavalier-Smith 1978, 1985)
C. General functional effects (as adjuncts to specific functional effects) based on sequence constraints of low specificity
–spacer functions (might include pseudogenes that maintain a given distance between functional genes)
–regulation through features of base composition and sequence of the thermodynamics of DNA strand separation or of other physicochemical features with general effects on DNA structure and function (e.g., Blaisdell 1983, 1985)
–function of DNA packing. This includes the action of proteins that bind selectively to satellite DNAs (Blumenfeld et al. 1978; Strauss and Varshavsky 1984), notably certain nonhistone proteins that co-purify with satellite-containing chromatin (Brutlag 1980)

* The present list is incomplete and the functions are often not independent

One cannot sustain the argument that the observed regularities may simply be due to their spontaneous restoration after decay. As the structure of satellite DNA shows, spontaneous processes lead to irregularities within regularities of sequence combinations (Miklos 1985). Nucleosome packing appears to be too regular, regionally, for being maintained or recovered exclusively through spontaneous nonselected processes. Furthermore, as mentioned, Bernardi and Bernardi (1986, 1987) provide substantive arguments in favor of a selectionist interpretation of the coincidence between base composition at silent codon positions on the one hand and in sequences flanking the structural genes on the other.

Packaging of DNA does not demonstrate DNA function—beyond the general function of packaging. Probably no "functionless" DNA would be tolerated, at least no large amounts of it would be, unless it can be properly packaged. Packaging may be essential for appropriate transmission of DNA from cell to cell and for functions linked to development and differentiation. If so, all or essentially all DNA may be subject to certain constraints in sequence features that are essential for packaging. In this case very little place may be left in genomes at which one can determine the true neutral mutation rate. This is the reason why I stated earlier in this article: "There is no telling when and where, *if anywhere*, the peregrinations through the genome of the true neutral mutation rate will end."

Evolution might preserve the required sequence features either by a periodic or by a continuous redress of their progressive decay. Consider the first mode. If the function of special sequence features in the bulk of noncoding sequences is limited to packaging itself and most noncoding DNA is otherwise nonfunctional, we may witness the frequent regeneration of the reputedly nonfunctional DNA from other nonfunctional DNA that is properly

packageable, with an accompanying loss of no-longer-properly-packageable adulterated nonfunctional DNA. One would have to postulate that such a sequence of events is ever-recurring. It could hardly take place regularly without the intervention of natural selection at the stages of sequence elimination, sequence regeneration, or both. The alternative is an intervention of natural selection in a more or less continuous mode rather than by periodic spurts, namely in a mode that involves the substitution of individual bases or the slippage of short sequences (Tautz et al. 1986) rather than the replacement of whole sequence-motif sectors and compositional sectors. In either case, why should cells and organisms take on the load of negative selection for the benefit of junk? There may be an answer to this question, to be proposed in this paper.

Functional Density in DNA: General and Specific Functions

While the case for selectionist constraints in the bulk of noncoding DNA sequences seems to be good, the predominant contrary view notwithstanding, the constraints on most noncoding DNA are relatively weak in terms of individual nucleotides.

What kind of functions can be linked to relatively weak constraints? If we were talking about proteins rather than about DNA, we would say: weak constraints indicate the engagement of the amino acids in general functions, not in specific functions. By general functions may be designated those that are helper functions, aimed at creating the proper molecular conditions for the specific functions to be properly carried out (Zuckerkandl 1976a). Extrapolating from the situation found in the proteins one would be tempted to surmise that nucleotides that can be exchanged at rates that may not be far from the "true" neutral mutation rate (not to mention nucleotide positions that can be freely lost or gained, with their identity frequently unascertainable in the process) either are functionless or, at best, participate in a minor way in some general function.

In fact it should be envisioned that in DNA the freedom of individual nucleotide sites to carry any of the four bases not only does not necessarily denote functionlessness, but does not even imply that the functions of the regions of DNA where nucleotides behave in this fashion are necessarily merely general. In DNA the frequent substitution of individual nucleotides may well be compatible with certain functions that deserve to be characterized as specific. Table 1 lists specific and general functions of DNA.

The definition of functional density (Zuckerkandl 1976a)—the proportion of sites engaged in specific functions—can easily be applied to DNA sequences such as promoters, in which the nucleotides committed to the specific functions, like amino acids in proteins, are well defined and highly constrained.[2] Because specificity of functions in proteins correlates with a high degree of residue invariance, most noncoding DNA would intuitively appear to have a very low functional density. I believe this inference not to be necessarily true, even though we are not yet able to pinpoint the involvement of variable nucleotide sites in specific functions and must therefore recognize that functional density, for most regions of DNA, is not at present determinable.

Estimating functional density of DNA or proteins obliges one not only to analyze in detail the ways in which functions are connected to their structural substrata, but also to sort out the commitments that the same sequence elements can, and often do, make simultaneously to different functions. Of interest to us here is not the quantitative estimate of functional density of DNA in the limited number of cases in which it can already be made, but rather the insights to be gained from the difficulties that arise when one tries to apply this concept to DNA. These difficulties are informative. They derive from the particular way in which DNA sequence elements appear to take charge of certain functions. Just as in proteins specific functions involve specific amino acids and specific amino acid sites, so in DNA specific functions would be expected consistently to involve specific nucleotides and nucleotide sites. In a fashion that is at variance with lessons learned from the proteins, nucleotides appear in fact to conform to certain functional imperatives by *combining looseness of sequence with specificity of effect.*

---

[2] The procedure of saturation mutagenesis, applied to the promoter of the mouse β-major globin gene (Myers et al. 1986) permits one to calculate the functional density (f.d.) of this promoter. From positions −101 to −1 (cap site not included), i.e., over 100 nucleotide positions, there are about 27 positions at which a substitution leads to a significant change in rate of transcription. For this segment, then, f.d. is about 0.27, which is about one-half or less than one-half of the f.d. of human hemoglobin polypeptide chains, in terms of the amino acids (0.52, miminum value; Zuckerkandl 1976a). In terms of the nucleotides in the coding sequences for these chains, f.d. is only 0.35, or a little higher when methionine and tryptophane are taken into account, for which the third codon position is coding. When, for more precision, the estimate is based on the nucleotide sequence of the human β-globin DNA sequence, the corresponding figure is only 0.31, as calculated by Dr. Takashi Gojobori (personal communication), and 0.32 taking into account met and trp. (A further increase in the polynucleotide f.d. would result from weighting the contribution of amino acids for which the third codon position is coding in terms of the alternative purine or pyrimidine.) Thus the f.d. of one of the noncoding sequences that controls the transcription of the β-globin gene, the promoter, is close to that of the structural gene itself. For specific functions, coding and noncoding sequences can use a similar fraction of the nucleotides.

In proteins, practically all sites are either engaged in specific functions, or in general functions, or in both. As an example, specific functions of a hemoglobin molecule are the binding of heme and oxygen, the binding of partner chains, the binding of the proton that controls the Bohr effect, that of 2,3-diphosphoglycerate, of carbon dioxide, of haptoglobin. Examples of general functions are ratios of polar to apolar amino acids, charge distribution and net charge, solubility, stabilization of α-helices, etc. In proteins, scarcely any sites can ever be considered as functionless. It must be assumed that specific function sites, in proteins, also contribute to general functions, but that the specific function constraints overrule general function constraints to the extent possible. When the imperatives of specific functions tend to interfere with the imperatives of general functions, general function sites presumably must make up for the molecular infirmity that would be created by the specific functions alone.

When it comes to applying the distinction between specific and general functions, the situation in DNA is found to be quite different from what it is in proteins. First, proteins or polypeptide chains are unambiguously defined molecular units. Within DNA, molecular units are not uniquely defined. There are a number of different ways, ideally, of apportioning molecular units, or units of molecular action, over a given region of DNA, all valid from one point of view or another.[3]

Second, in polypeptides general functions represent the physico-chemical bases that permit the macromolecules to carry out their specific functions. General functions do not exist independently of the specific functions that they serve. In polynucleotides there are few general functions in this sense. There are for instance spacer functions of DNA and perhaps the ability—if it is a function—of certain sequence characters to determine bends in the DNA (Hagerman 1986), etc. The classification of functions of DNA into specific and general given in Table 1 is more problematic than is a corresponding classification in the case of the proteins (Zuckerkandl 1976a). Moreover the functions listed are often not independent of one another.

The third aspect of the distinction between specific and general functions that does not seem to present special problems in proteins, but cannot be applied to DNA in the same straightforward way, hinges upon the fact that in proteins specific functions are linked to particular sets of amino acid residues and residue sites, whereas general functions are spread over sites in a way compatible with significant variability in the functional contribution of individual sites. The collective effects of a number of residues appear to be relatively fixed as far as the resultant overall value of a physico-chemical parameter is concerned, but the pathways by which these effects are achieved can be quite variable during the evolution of certain proteins, such as the hemoglobins. For instance, the same net charge of a protein can be obtained by a large number of combinations of different sequence characters. The number of functionally tolerable combinations is much reduced, however, by the fact that each residue participates simultaneously in several general functions and, in addition, may have a variable, indirect effect on the specific functions. If the latter circumstance is put in parenthesis, it may be said that sequence elements at the sites of general functions apparently count individually only inasmuch as they contribute to an overall effect obtained more or less interchangeably by their different combinations. On the other hand, specific functions in proteins are linked exclusively to sites whose identity and occupancy are heavily constrained, even though often not totally invariant in evolution.

In the case of certain specific DNA functions (Table 1), the individual "residues" engaged in them, the bases, also appear to have a great latitude in sequence arrangements. How could regions of DNA of high sequence variability carry out specific functions? The answer is clear enough when the specific function appears to depend exclusively on the quantity of nucleotides involved, independently of their nature and sequence (Cavalier-Smith 1978, 1985). When, on the other hand, specificity of function is linked to that of molecular interactions, it correlates with high affinity constants, which in turn seem to imply high sequence specificities. Again, in this case, how can high functional specificity be compatible with low sequence specificity?

Reconciling Specific Polynucleotide Function with Variable Sequence: Mass-Binding Proteins

According to a concept presented earlier (Zuckerkandl 1981) and since then also elaborated by Strauss and Varshavsky (1984), proteins that bind to periodically recurring sites in chromatin, even though they may bind with low affinity to their individual DNA receptor sites, may, through protein-protein-DNA multiple cooperative interactions, form very stable complexes and produce highly specific and

[3] There are, for example, sequence segments characterized by a high degree of specificity of action and a high degree of evolutionary conservation, such as promoters; functionally linked sets of such sequences, including 5′ and 3′ cis-regulatory sequences and the coding sequences; the unit formed by a set of introns and exons; the transcription unit; the relevant chromatin domain (chromatin loop), which frequently includes several transcription units (Scheer et al. 1976; Spring and Franke 1981). There might in certain cases be an additional functional unit between the latter two (Zuckerkandl 1981, p 154; Lawson et al. 1982).

topographically circumscribed effects. It may be that the establishment as well as the undoing or transformation of such complexes is involved directly in the specific functions of cellular determination.

The concept of multiple cooperative protein-DNA interactions could have relevance to the question of the function of DNA domains. Not only in metaphase, but in interphase as well, chromatin is organized into a series of loop-shaped domains that vary in length between ten or less and more than one hundred kilobase pairs (kb)—10 to 180 in the mouse according to Hancock and Boulikas (1982) (Cook and Brazell 1975; Benyajati and Worcel 1976; Paulson and Laemmli 1977; Igo-Kemenes and Zachau 1978; Marsden and Laemmli 1979; Razin et al. 1979; Zehnbauer and Vogelstein 1985). The loops are stabilized at their base by interactions with specific proteins (e.g., Lebkowski and Laemmli 1982) that form the so-called nuclear matrix or nuclear scaffold (see Lewis et al. 1984). One may presume, for instance, that the whole β-globin gene complex of man, plus an unspecified amount of sequences that flank the complex, are located in one loop (Stalder et al. 1980; Groudine et al. 1981).

In the retracted state, at least the larger loops must form a higher order structure not otherwise present in chromatin, and this structure is reasonably to be assumed to be stabilized by representatives of a class of regulator proteins that can be called mass-binding proteins. This name seems appropriate because these proteins bind to a relatively large number of sites along the DNA or along the RNA transcripts, in contradistinction to punctate-binding regulator proteins engaged in more highly specific interactions with one or a few receptor sites (Zuckerkandl 1981). Regulators binding to promoters are of the latter category, whereas histones are of the former.

Mass binding can be regulatory, as in the case of histone H1 (Thoma and Koller 1977; Schlissel and Brown 1984) and it is probably commonly, though not always, cooperative (Renz et al. 1978; Ruiz-Carillo et al. 1980). Mass-binding of nonhistone chromosomal proteins occurs; it can be sequence-specific (Strauss and Varshavsky 1984); the specificity involved is the attenuated specificity of sequence motifs; the attenuation of this local specificity does not preclude a high specificity of the overall interaction effect; and the interaction is cooperative. These features, predicted to be of general applicability in eukaryote genomes (Zuckerkandl 1981), with components of this picture proposed earlier (Zuckerkandl 1974, 1976b), have been verified in some cases, with the exception, as far as NHCPs are concerned, of widespread cooperativity. Cooperativity very likely obtains in the case of the α-protein (Strauss and Varshavsky 1984), though not in the cases of the related HMG-14 and HMG-17 (Mardian et al. 1980), nor HMG-1 and HMG-2 (Isackson et al. 1985).

It is well known that the condensation of eukaryotic chromatin is achieved through an extensive hierarchical folding of DNA, which is mediated by histones and nonhistone proteins (e.g., Ide et al. 1975; Sedat 1977; Georgiev et al. 1978). It seems likely that in most cases the folding and unfolding of chromatin are not mere effects, but functions, in relation to transcription, replication, or cell division. The mass-binding of certain types of proteins such as HMGs has been found to be instrumental in regional structural change (see Weisbrod 1982). In highly repetitive DNA, some NHCPs have been observed to replace quantitatively histone H1 (Musich et al. 1977, 1978). In another sequence environment, protein IP_{25}, upon induction of differentiation of mouse erythroleukemia cells, reaches 40% of histone H1 (Keppel et al. 1977). These proteins bind to nucleosome linker sequences, which directly or indirectly participate, at least in certain genes, in a sequence-determined phasing of the nucleosomes.

That the structure of *Drosophila* polytene chromosomes is maintained by mass-binding proteins has been suggested many years ago by the observation that the removal of the proteins by an alkaline urea solution conferred upon the polytene chromosomes an appearance of lampbrush chromosomes (e.g., Sorsa and Sorsa 1970; Bencze and Brasch 1979).

In polytene chromosomes of *Drosophila,* many bands display characteristic individual reactivities to ionically altered environments (Lezzi and Robert 1972; Kroeger and Müller 1973), no doubt expressing differences in DNA-protein interactions. These different reactivities are likely to reveal differences in DNA sequence characters and presumably also differences in the varieties of mass-binding proteins that may characterize different sets of bands.

Large-scale structural changes in DNA correlated with determination and differentiation have become apparent through the cytochemical work of Isidore and Eileen Gersh (Amenta et al. 1973). These results and their interpretation seem to have been largely forgotten. Typically for forgotten earlier work, some of the authors' conclusions were the same as those now drawn by molecular biologists. Recent studies of the action of certain endonucleases led to the rediscovery of the fact that determination and differentiation are accompanied by structural changes in chromatin and that chromatin passes through structural stages of preparation for transcriptional activity (Groudine and Weintraub 1982). There can be little doubt that the large-scale structural changes observed by the Gershes implicate mass-binding proteins, though DNP structure might in fact be partly determined by RNP (Simon et al. 1985).

There is evidence for the existence of proteins that bind to RNA transcripts in a way that can be characterized as mass binding (Macgregor 1980; Economidis and Pederson 1983; Risau et al. 1983). Some of these proteins display no specificity in regard to the origin and sequence structure of the primary transcripts. This does not imply that they may not preferentially bind to certain sequence motifs, as the histone octamers of nucleosomes appear to do in the case of at least some DNA. Other nonhistone chromosomal proteins are more selective. Thus a protein that binds to transcripts from a set of about ten lampbrush chromosome loops has been identified (Sommerville et al. 1978; Sommerville 1981). These original findings have been extended to further proteins and further sets of loops (Economidis and Pederson 1983; Lacroix et al. 1985). The studies of Beyer et al. (1980) show that the association of protein particles with nascent RNA varies characteristically from gene to gene in a manner that can be inferred to be polynucleotide sequence-dependent.

Thus certain protein-RNA interactions in the mass-binding (as well as punctate binding) mode may participate in putting constraints on noncoding DNA sequences. Admittedly, the functional relationships involved here are not clear. Again, what is function, what is mere effect? It has been relatively easy to promote the idea of large amounts of totally functionless DNA in the absence of any proof of function. On the other hand, asserting that the proteins that combine with RNA are functionless would be another matter. There are proteins with unknown function, but it would be implausible to conceive of functionless proteins. All proteins indeed are heavily constrained in most of their amino acid residues. The main function of some of the proteins that bind to RNA may well be linked to this very interaction. Proteins binding to RNA transcripts thus indirectly suggest a functional connection for the corresponding DNA itself.

In the noncoding parts of genomes, the structural substratum of function may largely be something more diffuse than we are accustomed to conceive. Such diffuseness does not imply weakness of a collective functional effect of a string of nucleotides. What we probably need to get accustomed to is the hierarchically different role played by individual nucleotides in different parts of genomes. With respect to the function of the majority of the noncoding sequences, the individual nucleotide appears to be a building stone of lower rank than it is in the coding sequences and in the rather short noncoding sequences endowed with the specific regulatory roles of "punctate" regulation. *To the majority of noncoding sequences whole DNA segments may be what individual nucleotides are to the coding sequences.*

Were much of the noncoding part of the genome to participate, if only indirectly, in specific functions of DNA, how could it be that noncoding DNA sequences can be gained and lost and be shifted around to the extent to which they are (e.g., Lewin 1981; Martin et al. 1983)? Such observations seem to contradict any involvement with specific function of the sequences that so behave. To resolve this matter, the particularities of noncoding DNA as a substratum for certain functions should be further clarified.

DNA Function and Functional Compatibility

It is noteworthy that a large majority of books that deal with biological functions refrain from stating what function is. One cannot take its meaning for granted, however, when one inquires about the functional status of most noncoding DNA. In general terms one may consider that a function is any activity of a component of a system when that activity is part of the coherent operation of the system. More specifically, one may define a biological function as any effect that participates in the organism's capacity to respond to, to exercise control over, and to gain or maintain independence from, the environment, all features that are expressed in the survival of the organism and the species *and* are established by positive or conserved by negative (stabilizing) natural selection.

One might infer from this definition that any feature of DNA that is positively selected or protected by negative selection is clearly functional and any feature that is not so selected, nonfunctional. However, nonselected features can be "hitchhiking" with selected ones. Moreover, in the case of DNA things are further complicated because of the apparent existence of distinct modes of functionality. With respect to function, four categories of DNA sequences need to be distinguished: sequences endowed with a pivotal function; sequences exerting detectable effects without filling a function, but compatible with functions carried out by closely-linked sectors of DNA; sequences without either function or effect, yet functionally compatible like the preceding ones; and sequences without function that do not display compositional or sequence features characterizing functional compatibility in the region of DNA concerned.

While certain cases covered by these categories will be taken in hand by the "neutral theory" (Kimura 1983), selectionists, curiously, can also stake claims on some of the same DNA segments because of the findings of Bernardi and collaborators (1985, 1986, 1987). These findings can indeed be construed to imply that not only pivotal function, but probably also functional compatibility may be selected for (as suggested, though not demonstrated, by trends in

base composition encompassing whole chromosomal bands and whole genomes) and the elimination of functional compatibility may be selected against (as suggested by the evolutionary stability of base composition in large sectors of noncoding sequences).

The distinction between functional modes is best explained with the help of a hypothetical example. When different domains of chromatin form high order structures, including the highest order structure that they are capable of forming, one of maximum compaction, this process can be expected to occur through the binding of certain mass-binding proteins. Let us consider the possibility that the difference between compacted and noncompacted domains exists in interphase chromosomes, that the difference is of relevance to gene regulation, and that there exists a set of structural variants of chromosomal proteins that bind with different affinities to different sets of domains, as a function of different predominant characters of base composition and sequence motifs. Such a situation would imply compositional and sequence constraints involving large regions of noncoding DNA. An alternative hypothesis leads to essentially the same expectation in regard to evolutionary constraints on noncoding parts of the genome. Namely, one may suppose that there is in every cell a set of variants of mass-binding proteins with preferential affinities for RNA primary transcripts, the latter being characterized by certain predominant base compositions and noncoding sequence motifs.

Thus two alternatives are considered: Every cell contains variants of mass-binding proteins that bind with different affinities (1) to different sets of chromatin domains (loops) as a function of the domains' base composition, and/or (2) to different sets of primary RNA transcripts as a function of the transcripts' base composition. Both situations would result in compositional constraints imposed upon sectors of noncoding DNA and upon silent codon positions.

The insertion into noncoding DNA of sequences that do not conform to the local sequence characteristics may well, beyond a threshold of length of individual or summed insertions, lead to disruptions or potentially damaging destabilizations of the high order structures, whether we deal with deoxyribonucleoprotein structures or with ribonucleoprotein structures. We shall assume that this holds for at least one of the two types of polynucleotide. The concept is illustrated in Fig. 1. We shall postulate that it is functionally important to maintain the potential to form the highest order structure in certain chromatin loops or to maintain, temporarily, the higher order structure of the transcripts, or both. We shall say that the DNA sequence features that permit these structures to be formed, even though the sequences may be of low functional density, carry out a pivotal function. A function involving whole DNA domains or large parts of them relative to a much smaller fraction of coding sequences would imply and thus explain, albeit partially, the existence of a "c-value paradox" (Zuckerkandl 1981).[4]

In this situation, a sufficient percentage of the sequences of the loop must have the base composition of those sectors of DNA that the relevant mass-binding regulator proteins preferentially interact with, and the distribution over the loop of these compositional sectors must be adequate. The insertion into the loop of additional sequences of the proper base composition might under certain circumstances lead to additional stabilization of the loop or of the correspondingly larger RNP structure synthesized on the loop; or RNP may change in some other property depending on size or conformation. There will undoubtedly be a change in the nature or concentration of some products of the enzymatic hydrolysis of the primary RNA tran-

[4] There are really two distinct parts to the c-value paradox. One part of the paradox is presented by eukaryote genomes of minimal size (see Cavalier-Smith 1978), in which there always is also a large excess of noncoding over coding sequences. The other part of the paradox arises as, for each evolutionary grade, the c-value varies from species to species between this minimum and a maximum. The range of this variation is limited in reptiles, birds, and mammals. It is notoriously striking in urodele amphibians. Of particular relevance is the way the further excess of DNA is distributed over the genomes. It is apportioned over both individual chromomeres and individual loops as they appear in lampbrush chromosomes (Macgregor 1980; Callan 1981; Scheer and Sommerville 1982). In particular, there are additions of noncoding sequences to individual loops, which are reflected in increased lengths of primary transcripts (Lengyel and Penman 1975; Scheer and Sommerville 1982). The two parts of the c-value paradox are probably not explainable in the same terms. The excess of noncoding sequences in minimum size genomes (c-value paradox I) is, in my opinion, likely to be significantly linked to structural requirements imposed by genetic functions (Zuckerkandl 1981). The variable excesses *over* this excess, which one may refer to as c-value paradox II, may be largely explainable in terms of DNA acting on cellular functions by its mere bulk, namely primarily on cell size (Cavalier-Smith 1978, 1985). There may be other functional connections for "c-value paradox II." The idea that c-value paradox II might be generated by the spreading in the genome of selfish DNA is difficult to reconcile with Keyl's (1965) observations on band widths in polythene chromosomes of different subspecies of *Chironomus thummi*. Keyl found that the DNA content per chromomere varied in the simple ratios 2, 4, 8, or 16, the higher values being confined to the same subspecies. He predicted the existence of loops as units of replication (Zehnbauer and Vogelstein 1985) and explained the observed doublings in the DNA content of chromosome bands by a doubling of such loops. The mechanism that he proposed or related mechanisms (Hancock and Boulikas 1982) may well also apply to the urodele genomes (Gall 1963) and may largely account for the process leading to c-value paradox II. It has not been definitively settled where the driving force for such an evolutionary process lies.

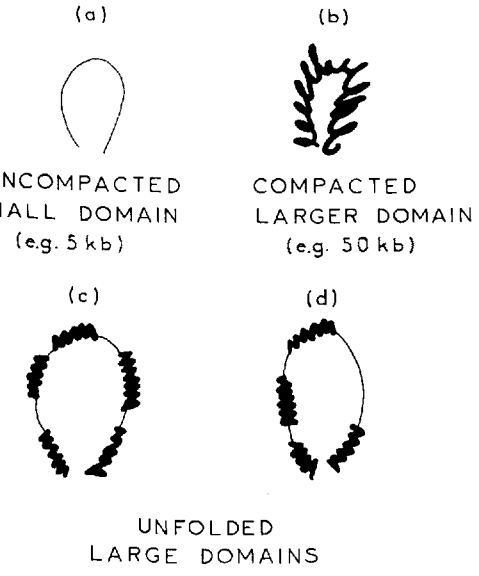

Fig. 1. Schematic illustration of the effects of domain size and of the distribution of sectors of different average base composition on the overall structure of DNP loops. (**a**), (**b**): the dependence of the highest order DNP structure on domain size. (a) Noncompacted, small domain, e.g., 5 kb (b) Compacted larger domain, e.g., 50 kb. (c), (**d**): effects of different distributions over domains of sectors of different average base composition. (~~~) sectors of given average base composition. (—) sectors of different average base composition. The domains are represented in the same relatively extended state as in (a). In (c), the sectors represented by the wiggled line predominate. Sectors of different base compositions are short enough so as not to destabilize substantially the higher order structure defined by the interaction between a specific protein and the compositional sectors represented by the wiggled line. (d) A long stretch of DNA of different average base composition is included in the domain. It is assumed that in this case the highest order DNP structure is destabilized

script. A change in *concentration* of some RNAs could thus be dependent upon a change in *length* of repetitive DNA. Such changes, again, are *effects*—the inserted sequence would not be ineffectual. Yet the effects need not be functional. (Though mere effects can eventually be turned into functions, we suppose that this has not occurred.)

In regard to DNP and perhaps even to RNP structure, relatively short additional transcribed sequences, and certainly short additional nontranscribed sequences, may in fact have no detectable effects. Thus additional sequence insertions may take place that neither exert a function nor produce an effect. This is when there is no ambiguity in characterizing the additional insert as functionally neutral. Still, in order to be so, we here suppose that the additional insert must by its composition, if not by certain sequence features, be functionally compatible with other regions of the loop. Base composition therefore is not functionally neutral in this functionally neutral insert. The insert remains neutral only as long as the base composition remains functionally adapted, "compatible."

Now let deletions occur. Let us assume that, in the case considered, the exact position of the deletion is not critical, but the amount is. This is a reasonable assumption, since there must be a critical loop size on either side of which the highest order structures are different. To be stabilized, each order of DNA structure indeed requires a minimum length of DNA, and this minimum increases with the order of the structure. Up to a certain extensiveness, the deletion will not compromise the function of the noncoding sequences in the area, in this case the potential stability of the loop's highest order structure. Beyond this point—not a very sharp boundary, one may presume—function will be impaired. Similarly, in the case of large RNP particles, it is possible that the length of the RNA transcript may be decreased by a certain amount without a qualitative change occurring, but if the transcript is further shortened, a radically new structural and functional situation may be created, conceivably even in regard to messages emitted by the cell, if parts of the RNA transcripts filled such a function. A quantitative change in such messages, coded for by repetitive DNA, could have repercussions for the organismal phenotype, and would represent a change in a specific function.

It is not possible to impart the attribute of pivotal functionality to any particular part of the sequence subject to deletions in our example, nor can the quality of mere functional compatibility be localized in any other part. Pivotal function, functional compatibility, and functional neutrality as well jointly pervade the large DNA segment that we suppose to be the target of deletions. It is clear that the functionalists and the neutralists have no reason ever to end their dispute over who this DNA segment belongs to. (It must be stressed that the neutrality referred to here is functional neutrality; whether or not a DNA insert or its deletion are neutral in the sense of spreading in a population by random drift is a matter that is not affected by the present discussion.) The controversy between the functionalists and the functional neutralists would perhaps die down if it were realized that, in appearance paradoxically, but according to the logic of special properties of noncoding DNA, *various modes of functionality and nonfunctionality can simultaneously take hold in a diffuse fashion of the same sector of DNA.*

When an increasingly extending deletion has reached the threshold above which function is impaired, then the remaining part of the original inserts, whichever part it happens to be, becomes

functionally pivotal. It is now a functionally pivotal noncoding sequence, even though it does not differ in any significant way from many other sequences. It has become pivotal by virtue of the fact that it now represents a sequence of minimum length. *"Junk" in DNA thus may not be inherent in sequences, but in their ratios to others.*

If the hypothetical example analyzed has enough realistic features, a straightforward, simple application of the notions of functional neutrality and nonneutrality is to be proscribed. These notions blur what appears to be a necessary distinction between pivotal functionality and functional compatibility. Special features of the noncoding parts of eukaryote genomes seem to oblige us to face up to this complication.

Evolutionary conservatism of sequence motifs or local base composition may be mistaken for a sign of effective or indispensable functionality, when it really is only a manifestation of conserved functional compatibility.

Conversely, the absence of evolutionary conservatism may mistakenly be taken as an indication of nonfunctionality. This is, for example, the case of the very large quantitative variations, from species to species, of repeated sequences, especially of the highly repeated simple sequences (Miklos 1985). Such variations in the amounts of any given type of satellite have nourished the impression that satellite DNA is functionless. From a functional point of view, however, the primary structure of such sequences is likely to be much more important than their quantity, beyond a certain threshold of quantity. Within some, probably imprecise, lower and upper boundaries, the quantities (and primarily they) of such sequences may be functionally neutral characters.

We have referred here to structure-function relationships in long stretches of noncoding DNA along which the conservation of function does not in general require a conservation of individual nucleotides. Such relationships apparently can also be found in much shorter noncoding regions in which specific functions are linked to a somewhat more stringent conservation of individual nucleotides. Thus the enhancer region of human cytomegalovirus appears to contain redundant sequence motifs that are all functional, but can be individually dispensed with. "Enhancers seem to be generally forgiving of all kinds of sequence manipulations" (Serfling et al. 1985). We seem to deal here with an excess of functional polynucleotide sequences that all conserve their functionality, presumably through some negative (purifying) selection, and yet are not all required for function. Each in turn probably could become functionally critical (pivotal), if the size of the enhancer were reduced to its minimum.

Is Interspersed Repetitive DNA Polite?

Short and long interspersed repeated DNA sequences (SINEs and LINEs), along with highly repeated simple-sequence heterochromatic DNA, together represent significant fractions of the genomes of higher eukaryotes. These sequences, especially SINEs and LINEs, are widely distributed through the genomes of higher eukaryotes (Singer 1982; Miklos 1985). They are found to differ considerably in numbers and distribution, as well as sometimes in kind, between even closely related species. Can one reasonably expect sequences that have been inserted into genomes at a large number of sites in recent evolutionary times to conform to local conditions of base composition and thus ensure that much of the moderately to highly repeated DNA, which encompasses selfish (Doolittle and Sapienza 1980; Orgel and Crick 1980) and reputedly ignorant (Dover 1980) DNA, will in fact prove to be *polite*? By polite, I mean respectful of general constraints and of functions carried out regionally by other DNA. Heterochromatin seems to form a base-compositional environment of its own and contains few genes (Hilliker and Appels 1980). The question therefore is of concern especially with respect to SINEs and LINEs. SINEs are only of the order of 0.3 kb long, which may be small enough for them, even if they lack "civic sense," not to disrupt significantly any compositional neighborhood. Furthermore, it is not excluded as yet that they fill a pivotal function (Schmid and Shen 1985), and they may well interact specifically with a nucleosome-phasing protein (Strauss and Varshavsky 1984). The length of LINEs, on the other hand, is of the order of 5 kb. There can be several types of LINEs per genome, and each sequence type can be represented about 10^4 to 10^5 times (Singer 1982).[5] LINEs therefore represent a challenge to the concept of polite DNA. Data, so far, are scarce. The concept is, however, supported by the findings of Meunier-Rotival et al. (1982), who studied the major family of LINEs in the mouse. Remarkably, not only are these repeat sequences confined to particular DNA components of the mouse, the two light major components, but *the compositional environment represented by the LINEs' flanking sequences matches the composition of the interspersed repeat.* These LINEs, therefore, clearly are polite DNA.

[5] That at least individual members of LINE families or sequences within LINE elements can have highly specific gene regulatory functions has been established by the discovery of the silencer function (Laimins et al. 1986). LINEs can also play the role of transforming elements, probably as a consequence of structural alterations of functional regions that they include (Cooper et al. 1986; Rogers 1986).

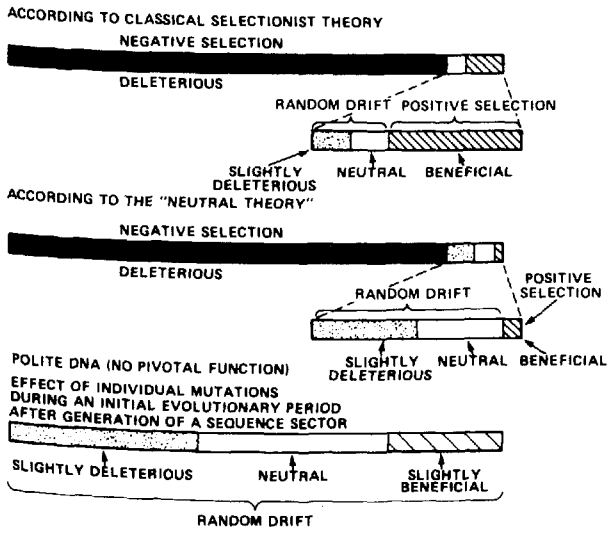

Fig. 2. Alternative views on the partitioning of mutations into deleterious and nondeleterious. The partitions represented do not aim at quantitative accuracy. Quantitative relationships are bound to vary as a function of the sequences considered and of environmental change. The selectionist and neutral theories are in agreement in regard to the predominant share, among mutations, of those subject to negative selection. They differ notably in regard to the share of mutations subject to positive selection. In the case of "polite" DNA, during an initial period after the establishment in the genome of a polite DNA component, mutations subject to negative selection presumably do not occur. An important fraction of the mutations may be slightly deleterious and as such become fixed in populations by random drift like neutral mutations. Since many mutations are expected to reverse the effects of the slightly deleterious ones, the contribution of slightly beneficial mutations, again behaving like neutral mutations, is likely to be more important than neutral mutation theory commonly envisions

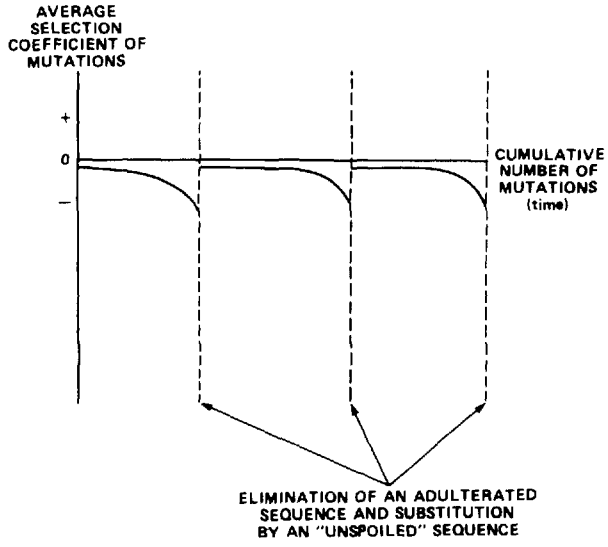

Fig. 3. Presumed cumulative effect of mutations on regions of "polite" DNA. After an initial evolutionary period during which mutations are nearly neutral (as indicated in the lower part of Fig. 2), the selection coefficient of further mutations becomes increasingly more negative. Eventually the whole sector of DNA presumably is eliminated through an event of negative selection. The adulterated sector may be replaced through random drift or positive selection by a newly inserted sector that displays the appropriate compositional features.

LINEs are candidates for selfish DNA. In regard to selfish DNA, the question generally asked is not whether such DNA exists, which seems practically unavoidable, given the existence of the apparatus that it requires, but to what extent. We propose here that much or most selfish DNA will be acceptable to genomes only if compatible with specific or general functions of host DNA and that such functions imply certain features of base composition and probably sequence motifs. The hypothesis is that, in any given compositional environment of the genome, selfish DNA frequently will either have to go or to prove not to be totally selfish by manifesting some consideration for the host genome. One may presume that much, if not most DNA is polite DNA. It may be required to lift its hat to its sequence neighbors.

Politeness and the "Neutral Theory"

The possibility that sequence constraints are widespread in noncoding sequences has often been negated a priori on the grounds that the species would be subject to excessive genetic load if selection had to handle a significant proportion of those mutations that occur in the largest component of the eukaryote genome (Kimura 1968, 1983). This argument has been controversial (e.g., Maynard Smith 1968), but persistent. Our only purpose here is to point out that Nature may resort to at least two dispositions capable of taking the sting out of the argument.

One of these dispositions is now demonstrated: the extensive use of sequence motifs rather than of rigorously defined sequences for protein-DNA interactions whose overall effect is structurally and, I predict, also functionally highly specific. A motif may accept a certain number of mutations as nearly neutral, up to a limit of divergence from an (abstract) consensus sequence. Thus, at some positions of the reiterated sequence motif, nucleotides may be replaced by less appropriate nucleotides through random drift. At other positions, however, or at the very same, chance may restore the appropriate nucleotides that had been lost. Chance, while adulterating the motif, may in this manner also contribute to keeping it in existence. Positive or negative selection will have to intervene less frequently than they would if rigorously defined sequences had to be evolved or maintained. Slightly beneficial mutations may therefore be much more frequent than the neutralists have so far considered, and they may spread by random drift just as the slightly deleterious mutations are supposed to do (Fig. 2).

The second process that would diminish genetic load is the following: In certain parts of the genome, notably in zones of highly repetitive sequences, mutations may be freely accepted as neutral until the sequence adulteration of a larger segment passes a certain threshold. At that time the adulterated sequence may be eliminated by negative selection and may be substituted by a better sequence, one regenerated by amplification from an appropriate master sequence. I have called this the forward creep-back leap mode of sequence evolution (Zuckerkandl 1975). It obviously would reduce radically the proportion of events of positive or negative selection necessary to maintain the sequence motifs and keep them at their proper places (Fig. 3).

Thus, under certain circumstances, the conservation of functional specificity in noncoding sequences is probably compatible with a relatively large dose of random drift, even though selection, at some stages, has to be decisively involved. In most noncoding regions, instead of intervening at the level of individual base substitutions, selection may affect whole DNA segments, once in an evolutionary while.

Conclusions

Polynucleotides appear to conform to certain functional imperatives by combining looseness of sequence with specificity of effect. In DNA, the freedom of a given nucleotide site to carry any of the four bases can be compatible with the participation of the nucleotide site and of its neighbors not only in a general function, but in principle even in a specific function. The latter situation presumably can occur thanks to cooperative mass-binding of certain proteins to sequence motifs that recur periodically along relatively long stretches of DNA. Regions of noncoding DNA that are relatively highly variable therefore need not be regions of low functional density. Low binding specificity is in principle compatible with highly specific *collective* regulatory effects (Zuckerkandl 1981; Strauss and Varshavsky 1984; Villet and Zuckerkandl, submitted for publication).

Periodic patterns of mostly as yet ill-defined sequence characters have been noted in *Drosophila* to occur in intergenic noncoding sequences as well as in larger intervening sequences (Keene and Elgin 1984), in keeping with the prediction that a general function of intervening sequences may be to participate in events of mass-binding of proteins that presumably lead to the formation of stable high order DNP structures. Evolutionary conservation in noncoding sequences of periodicities in sequence characters under relaxed sequence constraints may well carry the hallmark of natural selection in spite of the relaxed sequence constraints.

In a number of instances function may not be lodged specifically in any particular stretch of DNA of a compositional type. Function may be pervasive in such stretches, and may be assumed to require, regionally, a minimum quantity of sequence motifs of a kind. When the compositional sector, thanks to duplication events, displays more than this minimum amount of sequences, the extra-sequences are functionally superfluous; yet the anticipation is that it will not be possible to distinguish between sequences of the type that are functional from those that, because redundant, are in a sense functionless. Within a given region of DNA and in regard to certain functions the nonfunctionality implied by an excess of sequences is collective, and so is the functionality. Local additional sequences beyond the required minimum presumably have to conform to the compositional or sequence-motif patterns that characterize the region. If they do not, the function of the whole DNA region may be interfered with. Thus, even the superfluous components of a compositional or sequence-motif repeat pattern exercise a function of a kind, in a permissive mode. Within a zone of DNA, subsequences of significant length probably have to acknowledge the existence of the other regional sequences and conform to their compositional habit. Even additional, "superfluous" DNA thus may be subject to compositional constraints. Whether such additional and superfluous DNA is "selfish" or not, it seems that it cannot be "ignorant" of other DNA, but must be well-bred and "polite."

Pivotal function, functional compatibility and functional neutrality thus are expected jointly to pervade large noncoding DNA segments. During evolution, various modes of functionality and nonfunctionality can take hold of the same sectors of DNA. "Junk" in DNA may never be total junk in that it is, at least, properly organized junk. Furthermore, its character of junk may not be inherent in certain sequences, but rather in their quantitative relationship to others. Noncoding sequences that are merely compatible with the functions of neighboring sequences and are dispensable suggest an evolutionary mode in which highly deleterious or highly advantageous nucleotide substitutions do not occur. In such DNA zones, mutations probably are partitioned into slightly deleterious, neutral, and slightly advantageous. Random genetic drift thus appears capable of occupying the quasi totality of the scene. However, according to the model, this is only temporarily so. As the sequence sector under consideration decays with respect to the constraints imposed upon it, further mutations are expected often to be endowed with stronger negative selection coef-

ficients. Eventually, via a double event in which selection is likely to intervene, the adulterated sequence would be lost and sometimes (not necessarily) replaced by a more compatible, more "polite" sequence. This would slow the apparent overall rate of evolution of these noncoding sequences, which may nevertheless remain relatively rapid. The process is among those whereby different sequence constraints might be maintained—and might be changed—in different noncoding parts of the genome without generating an excessive genetic load. The genetic load argument therefore may not be a good reason for rejecting the notion that rapidly evolving DNA does fill some functional roles.

That such roles have not so far been convincingly demonstrated may be attributable to a scarcity in relevant technological advances. The nature of very high chromatin structures, their spatial distribution over the interphase genome, the developmental variations in this distribution, and their correlations with particular types of genes remain to be elucidated. Strides have been made in exploring the correlation between structural changes in chromatin and gene expression (e.g., Weisbrod 1982), but the nature of the structural changes is insufficiently understood. However, the lines of evidence referred to here do suggest the likelihood that rapidly evolving noncoding DNA does carry out functions of some kind. The accumulating evidence may now prompt a more widespread interest in addressing the germane structure-function problems.

Acknowledgments. I thank Drs. Ford Doolittle and Giorgio Bernardi for suggestions resulting from their critical reading of this manuscript, and Dr. Takashi Gojobori for comments and for kindly checking out a calculation. My appreciation, also, to Mrs. Ruth Reynolds; Mrs. Constance Canevari; and Mrs. Sandra Schwoebel for their exceptional bibliographical and secretarial contributions. This work was supported in part by The Japan Shipbuilding Industry Foundation.

References

Amenta PS, Gersh I, Gersh E (1973) Persistence of individuality of chromosomes during interphase, and the role of the nuclear membrane. In: Gersh I (ed) Submicroscopic cytochemistry. I. Proteins and nucleic acids, vol 1. Academic Press, New York, pp 365–375

Barnard EA, Cohen MS, Gold MH, Kim J-K (1972) Evolution of ribonuclease in relation to polypeptide folding mechanisms. Nature 240:395–398

Barrie PA, Jeffreys AJ, Scott AF (1981) Evolution of the β-globin gene cluster in man and the primates. J Mol Biol 149: 319–336

Bencze JL, Brasch K (1979) The morphology of normal and denatured polytene chromosomes from *Drosophila melanogaster*. Cytobios 25:93–104

Benezra R, Cantor CR, Axel R (1986) Nucleosomes are phased along the mouse β-major globin gene in erythroid and nonerythroid cells. Cell 44:697–704

Benyajati C, Worcel A (1976) Isolation, characterization, and structure of the folded interphase genome of *Drosophila melanogaster*. Cell 9:393–407

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22:363–365

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Bernardi G, Bernardi G (1987) The human genome and its evolutionary context. Cold Spring Harbor Symp Quant Biol "1986" (in press)

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Beyer AL, Miller OL Jr, McKnight SL (1980) Ribonucleoprotein structure in nascent hnRNA is nonrandom and sequence-dependent. Cell 20:75–84

Blaisdell BE (1983) Choice of base at silent codon site 3 is not selectively neutral in eucaryotic structural genes: It maintains excess short runs of weak and strong hydrogen bonding bases. J Mol Evol 19:226–236

Blaisdell BE (1985) Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. J Mol Evol 21:278–288

Blumenfeld M, Orf JW, Sina BJ, Kreber RA, Callahan, MA, Snyder LA (1978) Satellite DNA, H1 histone and heterochromatin in *Drosophila virilis*. Cold Spring Harbor Symp Quant Biol "1977" 42:273–275

Brutlag DL (1980) Molecular arrangement and evolution of heterochromatic DNA. Ann Rev Genet 14:121–144

Callan HG (1967) The organization of genetic units in chromosomes. J Cell Sci 2:1–7

Callan HG (1981) Lampbrush chromosomes. Proc Roy Soc Lond B214:417–448

Cavalier-Smith T (1978) Nuclear volume by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA c-value paradox. J Cell Sci 34:247–278

Cavalier-Smith T (1985) Cell volume and the evolution of eukaryotic genome size. In: Cavalier-Smith T (ed) The evolution of genome size. John Wiley & Sons, New York, pp 105–184

Cockerill PN, Garrard WT (1986) Chromosomal loop anchorage of the kappa immunoglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites. Cell 44:273–282

Cook PR, Brazell IA (1975) Supercoils in human DNA. J Cell Sci 19:261–279

Cooper GM, Goubin G, Diamond A, Neiman P (1986) Relationship of *Blym* genes to repeated sequences. Nature 320: 579–580

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603

Dover G (1980) Ignorant DNA? Nature 285:618–620

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Economidis I, Pederson T (1983) In vitro assembly of a premessenger ribonucleoprotein. Proc Natl Acad Sci USA 80: 4296–4300

Fry K, Salser W (1977) Nucleotide sequences of HS-α satellite DNA from kangaroo rat *Dipodomys ordii* and characterization of similar sequences in other rodents. Cell 12:1069–1084

Gall JG (1963) Chromosomes and cytodifferentiation. In: Locke M (ed) Cytodifferentiation and macromolecular synthesis. Academic Press, New York, pp 119–143

Georgiev GP, Nedospasov SA, Bakayev VV (1978) Supranucleosomal levels of chromatin organization. In: Busch H (ed) The cell nucleus, vol 6. Academic Press, New York, pp 1–34

Goldberg GI, Collier I, Cassel A (1983) Specific DNA sequences

associated with the nuclear matrix in synchronized mouse 3T3 cells. Proc Natl Acad Sci USA 80:6887–6891

Grantham R, Gautier C, Gouy M (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res 8:1893–1912

Groudine M, Weintraub H (1982) Propagation of globin DNAase I-hypersensitive sites in absence of factors required for induction: a possible mechanism for determination. Cell 30: 131–139

Groudine M, Peretz M, Weintraub H (1981) The structure and expression of globin chromatin during hematopoiesis in the chicken embryo. In: Stamatoyannopoulos G, Nienhuis AW (eds) Organization and expression of globin genes. Alan R. Liss, New York, pp 163–173

Hagerman PJ (1986) Sequence-directed curvature of DNA. Nature 321:449–450

Hancock R, Boulikas T (1982) Functional organization in the nucleus. In: International review of cytology, vol 79. Academic Press, New York, pp 165–215

Hilliker AJ, Appels R (1980) The genetic analysis of *D. melanogaster* heterochromatin. Cell 21:607–619

Holmquist GP, Dancis B (1979) Telomere replication, kinetochore organizers, and satellite DNA evolution. Proc Natl Acad Sci USA 76:4566–4570

Hutchison N, Weintraub H (1985) Localization of DNAse I-sensitive sequences to specific regions of interphase nuclei. Cell 43:471–482

Ide T, Nakane M, Anzai K, Andoh T (1975) Supercoiled DNA folded by non-histone proteins in cultured mammalian cells. Nature 258:445–447

Igo-Kemenes T, Zachau HG (1978) Domains in chromatin structure. Cold Spring Harbor Symp Quant Biol 1977 42: 109–118

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Isackson PJ, Cox DJ, Manning D, Reeck GR (1985) Studies on the interactions of HMG-1 and its homologs with DNA. In: Beckhor I (ed) Progress in nonhistone protein research, II. CRC Press, Boca Raton, Florida, pp 23–39

Jeffreys AJ (1982) Evolution of globin genes. In: Dover GA, Flavell RB (eds) Genome evolution. Academic Press, New York, pp 157–175

Jukes TH (1978) Codons and nearest-neighbor nucleotide pairs in mammalian messenger RNA. J Mol Evol 11:121–127

Keene MA, Elgin SCR (1984) Patterns of DNA structural polymorphism and their evolutionary implications. Cell 36:121–129

Keppel F, Allet B, Eisen H (1977) Appearance of a chromatin protein during the erythroid differentiation of Friend virus-transformed cells. Proc Natl Acad Sci USA 74:653–656

Keyl HG (1965) Duplikationen von Untereinheiten des chromosomalen DNS während der Evolution von *Chironomus thummi*. Chromosoma 17:139–180

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, Massachusetts

Koepsel RR, Kahn SA (1986) Static and initiator protein-enhanced bending of DNA at a replication origin. Science 233: 1316–1318

Kornberg R (1981) The location of nucleosomes in chromatin: specific or statistical? Nature 292:579–580

Kroeger H, Müller G (1973) Control of puffing activity in three chromosomal segments of explanted salivary gland cells of *Chironomus thummi* by variation in extracellular Na^+, K^+, and Mg^{2+}. Exp Cell Res 82:89–94

Lacroix JC, Azzouz R, Boucher D, Abbadie C, Pyne CK, Charlemagne J (1985) Monoclonal antibodies to lampbrush chromosome antigens of *Pleurodeles waltlii*. Chromosoma 92:69–80

Laimins L, Holmgren-Konig M, Khoury G (1986) Transcriptional "silencer" element in rat repetitive sequences associated with the rat insulin 1 gene locus. Proc Natl Acad Sci USA 83:3151–3155

Larsen A, Weintraub H (1982) An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin. Cell 29:609–622

Lawson GM, Knoll BJ, March CJ, Woo SLC, Tsai M-J, O'Malley BW (1982) Definition of 5′ and 3′ structural boundaries of the chromatin domain containing the ovalbumin multigene family. J Biol Chem 257:1501–1507

Lebkowski JS, Laemmli UK (1982) Non-histone proteins and long-range organization of HeLa interphase DNA. J Mol Biol 156:325–344

Lengyel J, Penman S (1975) RNA size and processing as related to different DNA content of two Dipterans: *Drosophila* and *Aedes*. Cell 5:281–290

Lewis DC, Lebkowski JS, Daly AK, Laemmli UK (1984) Interphase nuclear matrix and metaphase scaffolding structures. J Cell Sci Suppl 1:103–122

Lewin R (1981) Do jumping genes make evolutionary leaps? Science 213:634–636

Lezzi M, Robert M (1972) Chromosomes isolated from unfixed salivary glands of *Chironomus*. In: Beermann W (ed) Developmental studies on giant chromosomes. Springer-Verlag, New York, pp 35–57

Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature Lond 292:237–239

MacGregor HC (1980) Recent developments in the study of lampbrush chromosomes. Heredity 44:3–35

Madhani HD, Bohr VA, Hanawalt PC (1986) Differential DNA repair in transcriptionally active and inactive proto-oncogenes: c-*abl* and c-*mos*. Cell 45:417–423

Mardian JKW, Paton AE, Bunick GJ, Olins DE (1980) Nucleosome cores have two specific binding sites for nonhistone chromosomal proteins HMG 14 and HMG 17. Science 209: 1534–1536

Marsden MPF, Laemmli UK (1979) Metaphase chromosome structure: evidence for a radial loop model. Cell 17:849–858

Martin G, Wiernasz D, Schedl P (1983) Evolution of *Drosophila* repetitive-dispersed DNA. J Mol Evol 19:203–213

Maynard Smith S (1968) "Haldane's Dilemma" and the rate of evolution. Nature 219:1114–1116

Meunier-Rotival M, Soriano P, Cuny G, Strauss F, Bernardi G (1982) Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc Natl Acad Sci USA 79:355–359

Miklos GLG (1985) Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum Press, New York, pp 241–321

Miklos GLG, Gill AC (1981) The DNA sequences of cloned complex satellite DNAs from Hawaiian *Drosophila* and their bearing on satellite DNA sequence conservation. Chromosoma 82:409–427

Miyata T, Yasunaga T (1981) Rapidly evolving mouse α-globin-related pseudogene and its evolutionary history. Proc Natl Acad Sci USA 78: 450–453

Moreau J, Marcaud L, Maschat F, Kejzlarova-Lepesant J, Lepesant J-A, Scherrer K (1982) A + T-rich linkers define functional domains in eukaryotic DNA. Nature 295:260–262

Musich PR, Brown FL, Maio JJ (1977) Subunit structure of chromatin and the organization of eukaryotic highly repetitive DNA: nucleosomal proteins associated with a highly repetitive mammalian DNA. Proc Natl Acad Sci USA 74:3297–3301

Musich PR, Brown FL, Maio JJ (1978) Mammalian repetitive DNA and the subunit structure of chromatin. In: Symposia on quantitative biology 1977, vol 42 (2). Cold Spring Harbor Laboratory, pp 1147–1160

Myers RM, Tilly K, Maniatis T (1986) Fine structure genetic analysis of a β-globin promoter. Science 232:613–618

Ohno S (1972) So much "junk" data in our genome. In: Smith HH (ed) Evolution of genetic systems. Gordon-Breach, New York, pp 366–370

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607

Paulson JR, Laemmli UK (1977) The structure of histone-depleted metaphase chromosomes. Cell 12:817–828

Prunell A, Kornberg RD (1978) Relation of nucleosomes to DNA sequences. In: Symposia on quantitative biology 1977, vol 42 (1). Cold Spring Harbor Laboratory, pp 103–108

Razin SV, Mantieva VL, Georgiev GP (1979) The similarity of DNA sequences remaining bound to scaffold upon nuclease treatment of interphase nuclei and metaphase chromosomes. Nucleic Acids Res 7:1713–1735

Renz M, Nehls P, Hozier J (1978) Histone H1 involvement in the structure of the chromosome fiber. Cold Spring Harbor Symp Quant Biol 1977 42:245–252

Risau W, Symmons P, Saumweber H, Frasch M (1983) Non-packaging and packaging proteins of hnRNA in *Drosophila melanogaster*. Cell 33:529–541

Rogers J (1986) Relationship of *Blym* genes to repeated sequences. Nature 320:579

Ruiz-Carrillo A, Puigdomenech P, Eder G, Lurz R (1980) Stability and reversibility of higher order structure of interphase chromatin: Continuity of deoxyribonucleic acid is not required for maintenance of folded structure. Biochemistry 19: 2544–2554

Schaufele F, Gilmartin GM, Bannwarth W, Bernstiel ML (1986) Compensatory mutations suggest that base-pairing with a small nuclear RNA is required to form the 3′ end of H3 messenger RNA. Nature 323:777–781

Scheer U, Sommerville J (1982) Sizes of chromosome loops and hnRNA molecules in oocytes of amphibia of different genome sizes. Exp Cell Res 139:410–416

Scheer U, Franke WW, Trendelenburg MF, Spring H (1976) Classification of loops of lampbrush chromosomes according to the arrangement of transcriptional complexes. J Cell Sci 22:503–519

Schlissel MS, Brown DD (1984) The transcriptional regulation of Xenopus 5S RNA genes in chromatin: the roles of active stable transcription complexes and histone H1. Cell 37:903–913

Schmid CW, Shen C-KJ (1985) The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum Press, New York, pp 323–358

Sedat J (1978) A direct approach to the structure of eukaryotic chromosomes. Cold Spring Harbor Symp Quant Biol "1977" 42:331–350

Serfling E, Jasin M, Schaffner W (1985) Enhancers and eukaryotic gene transcription. Trends in Genetics 1:224–230

Simon JA, Sutton CA, Lobell RB, Glaser RL, Lis JT (1985) Determinants of heat shock-induced chromosome puffing. Cell 40:805–817

Singer MF (1982) SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28: 433–434

Sommerville J (1981) Immunolocalization and structural organization of nascent RNP. In: The cell nucleus, vol VIII. Academic Press, New York, pp 1–57

Sommerville J, Crichton C, Malcolm D (1978) Immunofluorescent localization of transcriptional activity on lampbrush chromosomes. Chromosoma 66:99–114

Sorsa V, Sorsa M (1970) Ultrastructure of induced transitions in the chromatin organization of *Drosophila* polytene chromosomes. Chromosoma 31:346–355

Spring H, Franke WW (1981) Transcriptionally active chromatin in loops of lampbrush chromosomes at physiological salt concentrations as revealed by electron microscopy of sections. Eur J Cell Biol 24:298–308

Stalder J, Larsen A, Engel JD, Dolan M, Groudine M, Weintraub H (1980) Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I. Cell 20:451–460

Strauss F, Varshavsky A (1984) A protein binds to a satellite DNA repeat at three specific sites that would be brought into mutual proximity by DNA folding in the nucleosome. Cell 37:889–901

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Thoma F, Koller Th (1977) Influence of histone H1 on chromatin structure. Cell 12:101–107

Udvardy A, Schedl P (1983) Structural polymorphism in DNA. J Mol Biol 166:159–181

Udvardy A, Schedl P, Sander M, Hsieh T-S (1985) Novel partitioning of DNA cleavage sites for *Drosophila* topoisomerase II. Cell 40:933–941

Vanin EF, Goldberg GI, Tucker PW, Smithies O (1980) A mouse α-globin-related pseudogene lacking intervening sequences. Nature 286:222–226

Weintraub H (1983) A dominant role for DNA secondary structure in forming hypersensitive structures in chromatin. Cell 32:1191–1203

Weintraub H (1985) Assembly and propagation of repressed and derepressed chromosomal states. Cell 42:705–711

Weisbrod S (1982) Active chromatin. Nature 297:289–295

Zehnbauer BA, Vogelstein B (1985) Supercoiled loops and the organization of replication and transcription in eukaryotes. BioEssays 2:52–54

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC (1980) Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc Natl Acad Sci USA 77:2158–2162

Zuckerkandl E (1965) Remarques sur l'évolution des polynucléotides comparée à celle des polypeptides. Bull Soc Chim Biol 47:1729–1730

Zuckerkandl E (1974) A possible role of "inert" heterochromatin in cell differentiation. Biochimie 56:937–954

Zuckerkandl E (1975) The appearance of new structures and functions in proteins during evolution. J Mol Evol 7:1–57

Zuckerkandl E (1976a) Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. J Mol Evol 7:167–183

Zuckerkandl E (1976b) Gene control in eukaryotes and the c-value paradox. J Mol Evol 9:73–104

Zuckerkandl E (1981) A general function of noncoding polynucleotide sequences. Molec Biol Rep 7:149–158

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166