Volume 16 Number 14 1988 Nucleic Acids Research

Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism

Roger D.Kornberg and Lubert Stryer

Department of Cell Biology, Stanford University School of Medicine, Fairchild Science Building, Stanford, CA 94305, USA

Received June 1, 1988

ABSTRACT

Expressions are derived for distributions of nucleosomes in chromatin. Nucleosomes are placed on DNA at the densities found in bulk chromatin, and their locations are allowed to vary at random. No further assumptions are required to simulate the periodic patterns of digestion obtained with various nucleases. The introduction of a boundary constraint, due for example to sequence-specific protein binding, results in an array of regularly spaced nucleosomes at nonrandom locations, similar to the arrays reported for some genes and other chromosomal regions.

INTRODUCTION

A paradox has arisen in studies of the location and distribution of nucleosomes. Nonrandom locations of nucleosomes have been reported for many DNA sequences in a variety of organisms (1-22). Such findings extend in some cases to nucleosomes assembled *in vitro* from purified histones and DNA (23-26). There is evidently a degree of sequence-specificity in the histone-DNA interaction. On the other hand, almost all DNA sequences in eukaryotes are packaged in arrays of nucleosomes, so the sequence-specificity cannot be too great. Moreover, arrays of nucleosomes are periodic, as revealed by ladders of bands in gels following nuclease digestion. The periodicity (spacing of nucleosomes) may vary among cell types of an organism, although the DNA sequence remains the same (27). The locations of nucleosomes must, then, vary among cell types as well. How can the evidence for specific locations of nucleosomes be reconciled with that for variation? A way of resolving the paradox was put forward (28) in which the location of nucleosomes varies at random on a DNA sequence but appears nonrandom in the average over many copies of the sequence. The sole requirement for such ordering of nucleosomes is the occurrence of a boundary at a point in the DNA sequence. The ordering, or apparent positioning, of nucleosomes then follows from their high density on DNA. Here we demonstrate the theoretical validity of this

"statistical positioning" idea. We calculate statistical distributions of nucleosomes with minimal assumptions and obtain good agreement with experiments.

The assumptions of the theory and the experimental results derive from studies of the nuclease digestion of chromatin (27). Digestion occurs preferentially in the linker region between nucleosomes. Brief digestion results in a series of DNA fragments whose lengths are multiples of a unit size, corresponding to the periodicity or repeat distance (spacing) of nucleosomes. The periodicity is typically about 200 bp, and ranges from about 165 bp in fungi to 240 bp in sea urchin sperm. Upon extended digestion with some enzymes, such as micrococcal nuclease, the unit length is reduced by trimming from the ends, with a pause at about 166 bp, until a limit is reached at about 146 bp. The 166 bp particle contains a full complement of histones, including one molecule of H1, while the 146 bp, or "core", particle has lost H1. In contrast with the wide range of periodicities (repeat lengths) of nucleosomes, the core length has been conserved in evolution. The variation in periodicity is due entirely to variation of the linker length between cores.

Locations of nucleosomes may be mapped by the "indirect end-labeling" technique (2,29). The procedure entails brief nuclease digestion of chromatin, followed by extraction of the DNA and limit digestion with a restriction enzyme. The resulting DNA fragments are resolved by gel electrophoresis and blot-hybridized with a probe from near a restriction site of interest. The lengths of DNA fragments revealed in this way measure distances from the restriction site to points of cutting by nuclease (linker regions) in the original chromatin. If nucleosomes are located at specific positions, then bands separated by at least 146 bp will be seen in the blot; if nucleosomes are located randomly, then there will be no bands, or else a band pattern similar to that obtained from digestion of naked DNA, reflecting the sequence preferences of the nuclease used.

Indirect end-labeling analysis has been performed for many regions containing functional elements of chromosomes, such as genes, centromeres, telomeres, and ARS elements (2-22). In most cases, band patterns with periodicities (but showing some deviation from a perfectly regular spacing) were observed. The periodicities corresponded with those of the bulk chromatin of the organism, and were therefore attributed to arrays of nucleosomes. The band patterns were interrupted by one or more sites unusually susceptible ("hypersensitive") or particularly resistant to digestion. Hypersensitive sites often contain enhancer and promoter sequences, and the appearance of such sites is

correlated with gene activation (30-39). Nuclease-resistant sites may contain special sequences as well, for example the conserved DNA elements of centromeres (6). Both hypersensitive and resistant sites have been shown to bind proteins that are sequence-specific and have high affinities for DNA (40-44). Besides revealing nonrandom locations of nucleosomes near specific protein binding sites, indirect end-labeling analysis has given evidence for randomly located nucleosomes. The band patterns obtained from some gene regions were essentially the same as those given by naked DNA (21,45). The gene regions in question were transcriptionally inactive, and showed no evidence of specific protein binding.

The correlation of nonrandom nucleosome locations with specific protein binding lends credence to the statistical positioning idea. Specific protein binding may create the boundaries required for statistical positioning by excluding nucleosomes from the binding sites. The occurrence of hypersensitive and nuclease-resistant regions attests to the formation of boundaries in this way. It has also been shown that sequence-specific transcription factors and nucleosomes compete for binding to a promoter (46), and that restriction enzymes fail to cut DNA in nucleosomes (15,47), indicative of mutual exclusion between nucleosomes and other proteins.

In the theoretical test of statistical positioning that follows, nucleosomes are not allowed to cross boundaries. They are treated as noninteracting, noninterpenetrating, and fixed in place on DNA. No further assumptions are required to simulate nuclease digestion experiments.

THEORY

Consider a very long array of nucleosomes (beads) on DNA. The array can be regarded as an ensemble of two kinds of objects, beads and bare sites. Let d denote the diameter of a bead (a number of base pairs, but a dimensionless quantity) and L the mean length of bare DNA (the mean linker length). The probability v of selecting a bead from this ensemble is $1/(1+L)$, and the probability u of selecting a bare site is $1-v$ or $L/(1+L)$. Note that u differs from the average probability that a site in the DNA is bare, given by $L/(d+L)$. The probability of selecting n beads and h bare sites is $v^n u^h$. They can be chosen in $(n+h)!/n!h!$ ways. Hence, w(g,n), the weighted number of ways, is given by

$$w(g,n) = v^n u^h \frac{(n+h)!}{n!h!} \qquad [1]$$

where g is the number of base pairs, g = h + nd. A domain of length g can contain any integral number of beads from 0 to n_{max}, where n_{max} is the largest integer less than g/d. Let s(g) be the sum of w(g,n) over all permissible values of n.

$$s(g) = \sum_{n=o}^{n_{max}} w(g,n) \qquad [2]$$

It can be shown that s(g) converges as g approaches infinity.

$$s_{\infty} = \lim_{g \to \infty} s(g) = \frac{1+L}{d+L} \qquad [3]$$

These expressions for w(g,n) and s(g) can be used to calculate the probability of any particular distribution of beads and bare sites. Consider the distribution on a DNA molecule of length z. The probability p(t) that a site t nucleotides from an end is bare is given by

$$p(t) = \frac{s(t-1)w(1,0)s(z-t)}{s(z)} \qquad [4]$$

This expression is derived by noting that a bare site partitions the DNA into three domains: a domain of length t-1 that can contain beads, the bare site itself, and a domain of length z-t that can contain beads. These domains give rise to the s(t-1), w(1,0), and s(z-t) terms in the numerator. The denominator s(z) represents all possible arrangements of beads and bare sites over an unpartitioned length of DNA.

Equation 4 can be used to calculate the probability of a bare site t nucleotides away from an end of a very long DNA molecule. The s(z-t) and s(z) terms are both replaced by s_{∞}, giving

$$p(t) = s(t-1)w(1,0) \qquad [5]$$

The derivation of Equation (4) can also be extended to the case of two bare sites a and b nucleotides from the end of the DNA. The probability p(a,b) that both sites are bare is given by

$$p(a,b) = \frac{s(a-1)\ w(1,0)\ s(b-a)\ w(1,0)\ s(z-b)}{s(z)} \qquad [6]$$

If z, the length of the DNA, is large, and sites a and b are far from the ends, then s(a-1) and s(z-b) are very nearly equal to s_∞, and the expression simplifies to

$$p(a,b) = w^2(1,0)\ s_\infty\ s(b-a) \qquad [7]$$

In what follows, we assume a probability c of cutting by nuclease at a bare site. A DNA fragment of length y=b-a will be produced if cuts occur at a and b but not at an intermediate position. The probability of cutting at two sites is c^2. The probability of not cutting at an intermediate site is taken into account by replacing s(y) with q(y), given by

$$q(y) = \sum_{n=0}^{n_{max}} w(y,n)\ (1-c)^h \qquad [8]$$

where h, the number of bare sites, is equal to y-nd. The $(1-c)^h$ term in this expression is the probability of not making a cut in the region between a and b. The probability r(y) of a fragment of length y resulting from a pair of cuts in the interior of a long DNA molecule is then

$$r(y) = c^2 w^2(1,0)\ s_\infty q(y) \qquad [9]$$

RESULTS

We have computed distributions of nucleosomes from the expressions derived above to simulate nuclease digestion experiments. The value of the bead diameter, d, was taken as 166 bp, since this length of DNA is protected from digestion by a complete nucleosome, including H1. Values of L, given by the nucleosome repeat (the periodicity) minus 166 bp, were approximately those determined experimentally for a range of organisms (27): 15 bp (HeLa cells); 30 bp (rat liver, chicken oviduct); 45 bp (chicken erythrocytes); 75 bp (sea urchin sperm). As a check on the theory and the computer programs, we calculated s(g) for large values of g from Equation 2. For L values ranging from 30 to 75 bp, s(5000) was within 0.001% of (1+L)/(d+L), as expected from Equation 3. For an L value of 15 bp, agreement was to within 0.6%.

The simplest case to simulate is the nuclease digestion of a region adjacent to a boundary, analyzed by indirect end-labeling. The reason is that only a

single cut by nuclease is involved. The probability of this cut at a particular site may be equated to the probability that the site is bare, given as a function of distance from the boundary by Equation 5. It is assumed that a single bare nucleotide between nucleosomes can be cut by nuclease. While the actual length of bare DNA required for nuclease action is not known, it makes little difference, since it enters primarily through the factor of w in Equation 5. (The argument of s also changes, but the numerical effect is negligible). The factor of w determines the amplitude of the probability distribution, but does not affect its shape. For a single bare site accessible to nuclease, and an L of 30, the value of w is given by

$$w(1,0) = u = 0.96$$

whereas if 10 bare nucleotides were required for nuclease action, then w would become

$$w(10,0) = u^{10} = 0.72$$

Probability distributions computed from Equation 5 showed a damped oscillation about the average probability that a site is bare (Figure 1). The period of the oscillation was nearly identical with the nucleosome repeat, d+L. The theory thus predicts a regular spacing of nucleosomes adjacent to the boundary, with the periodicity observed for bulk chromatin, although the distribution of nucleosomes was assumed to be random. The nonrandom locations of nucleosomes persisted for only a limited distance from the boundary, due to the damping effect. The damping was more pronounced for larger values of L, but even with a value of 75 bp, the largest that has been measured, an apparent ordering of nucleosomes was observed.

Having computed the effect of a single boundary on the nuclease digestion pattern revealed by indirect end-labeling, we wished to introduce a second boundary, defining a region of limited extent. Such regions occur in chromosomes, where boundaries formed by specific protein-binding sites may be closely spaced. The nuclease digestion patterns for regions of various size were computed from Equation 4. Again, damped oscillations were observed, with the same periodicity of d+L as that measured for nucleosomes in bulk chromatin, except for small regions whose sizes were nonintegral multiples of the

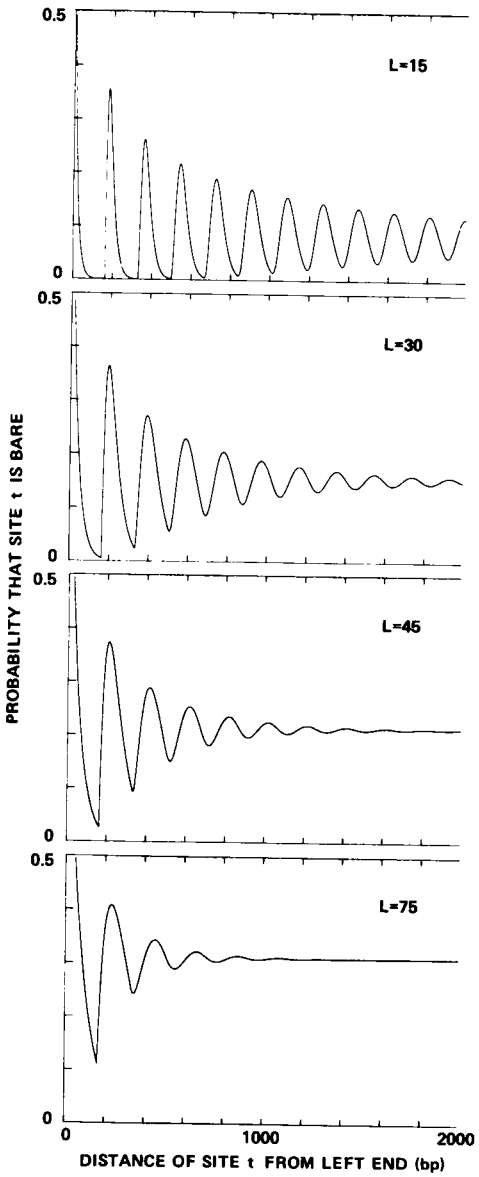

Figure 1. Simulation of indirect end-labeling experiments for regions bounded on the left side. Values of p(t) were computed from Equation 5 for the values of L indicated.

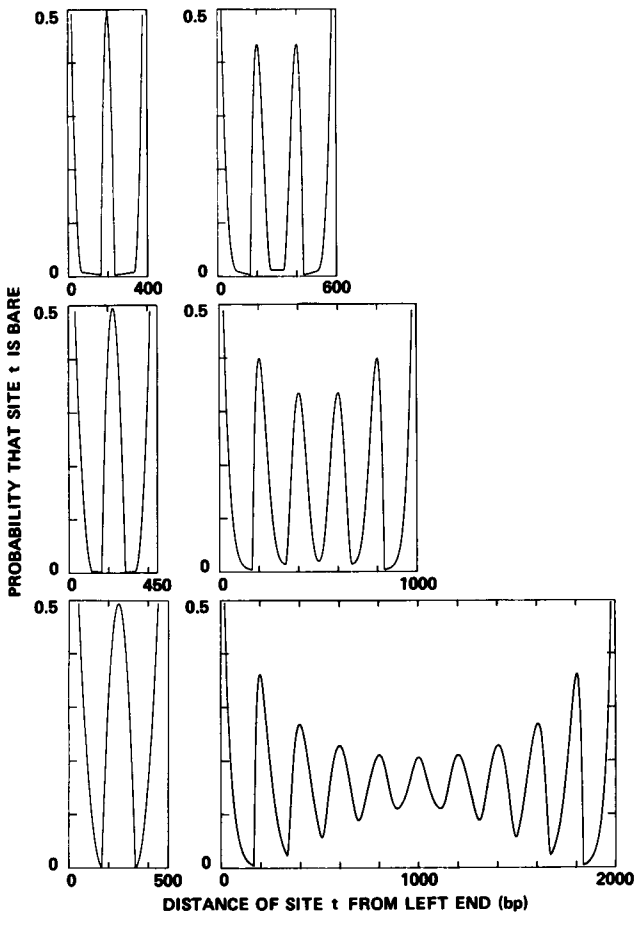

Figure 2. Simulation of indirect end-labeling experiments for regions bounded on both sides. Values of p(t) were computed from Equation 4 with L = 30 for various lengths z of the region.

periodicity (Figure 2). In such cases, the nucleosomes were evenly distributed across the region, with a spacing different from the periodicity in bulk chromatin (for example spacings of 225 and 250 bp, compared with a periodicity for bulk chromatin of 196 bp, in the first two plots of Figure 2). This effect was also seen upon varying L rather than the size of the region. The spacing of nucleosomes was 200 bp for all values of L between 15 and 60 bp in the example shown (Figure 3).

The simplest nuclease digestion experiment, in which the resulting DNA fragments are revealed directly rather than by indirect end-labeling, is the most

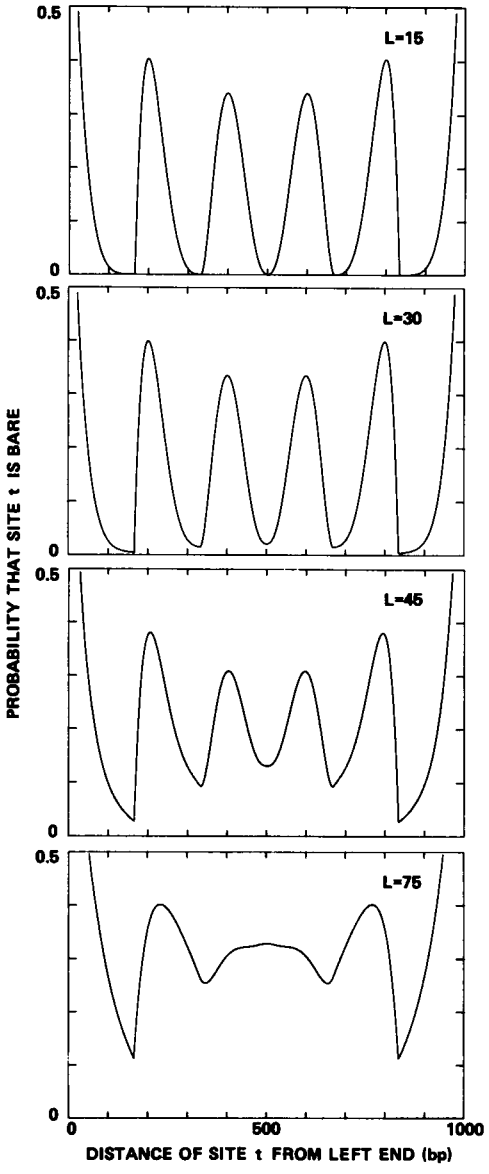

Figure 3. Simulation of indirect end-labeling experiments for regions bounded on both sides. Values of p(t) were computed from Equation 4 with z = 1000 for the values of L indicated.

difficult to simulate, since each fragment is produced by two cuts, and only fragment lengths, rather than locations of cuts, are measured. The probability of obtaining a fragment of length y is proportional to q(y), given by

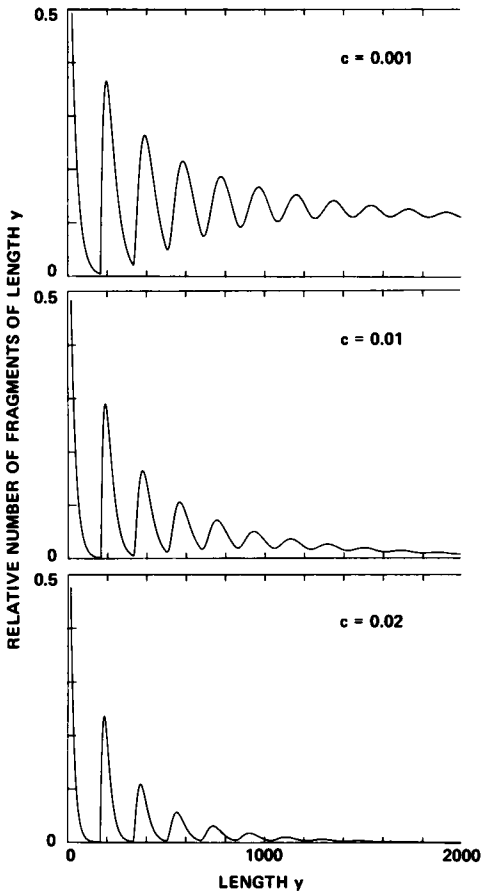

Figure 4. Simulation of a simple nuclease digestion experiment. Values of q(y) were computed from Equation 8 with L = 30 for the values of c indicated. Values of y are in bp.

Equation 8. For small values of the cutting frequency, c, Equation 8 reduces to Equation 5, aside from constant factors, and the results of the simple nuclease digestion experiment then have the same functional form as those of indirect end-labeling. As c approaches 1, q(y) becomes vanishingly small, except for y = nd. Computations of q(y) from Equation 8 illustrated this dependence of the nuclease digestion pattern on the cutting frequency (Figure 4). For c = 0.001, q(y) was indistinguishable from p(y) (compare top panel in Figure 4 with second panel in Figure 1). For c = 0.02, the oscilla-

tions of q were more sharply peaked and the peak positions were shifted towards smaller values of y.

DISCUSSION

The computations presented here yield ladders of bands with the characteristics expected from simple nuclease digestion experiments. The computations are also in good agreement with the results of indirect end-labeling experiments. Both the computations and the experimental results show nonrandom locations of nucleosomes adjacent to a boundary, with a regular spacing essentially the same as the periodicity of nucleosomes in bulk chromatin. The computations show a gradual loss of order with increasing distance from the boundary, and where this point has been pursued experimentally it has been observed as well. In two studies, the boundary and flanking sequences have been manipulated to obtain proof of the boundary effect. The first study (48) began with a region of a yeast plasmid showing little ordering of nucleosomes. When the 5'-end of a gene, with presumed binding sites of transcription factors, was introduced in the region, a pronounced ordering of nucleosomes resulted. The second study (21) concerned an array of nucleosomes found in nonrandom locations adjacent to a regulatory region in a yeast plasmid. When the sequences on either side of the regulatory region were changed, the locations of nucleosomes remained the same. It was evidently a feature of the regulatory region that determined the locations of nucleosomes. This feature was identified as a binding site of an abundant protein factor. To summarize, either changing a boundary and keeping the flanking sequences the same, or keeping the boundary constant and changing the flanking sequences, had the same outcome: the boundary was responsible for the ordering of nucleosomes.

Boundary effects are of course not the only influences that may cause nonrandom locations of nucleosomes. Preferential binding of the histone octamer to certain DNA sequences has been well documented (23-26), and the molecular basis of such preferences is at least partly understood (26). In addition, there may be ordering due to interactions between nucleosomes, and to the formation of higher order chromatin structures. It is useful, in considering these influences, to distinguish among various strengths, or "orders" of effect. The binding of a sequence-specific protein to DNA creates a boundary whose effect upon neighboring nucleosomes is of first order near the boundary and decays to lower order with increasing distance from the boundary. The preferential binding of histones to certain sequences is a second order effect, whose influence upon neighboring nucleosomes is then of third order.

The interplay between these various effects is apparent in some experimental data. For example, the patterns of nuclease digestion near hypersensitive sites often show deviations from regularity that may reflect local sequence preferences of the histones (or of cutting by the nuclease). In the analysis of nucleosome locations near a regulatory region in a yeast plasmid mentioned above, the boundary effect of the regulatory region extended for only about 500 bp. At greater distances, the locations of nucleosomes were characteristic of the particular DNA sequence used and independent of the boundary.

The nonrandom locations of nucleosomes resulting from boundary effects, sequence preferences, and so forth are unlikely to have biological significance. They lack the specificity needed for exposing or covering up regulatory elements or for other functional roles that have been proposed. The nonrandom locations computed here serve to illustrate the point. These locations are seen only in the average over many genomes. In a sizeable fraction of a population (given by the minima of the oscillations in Figures 1-3), nucleosomes will cover up the very sequences that are exposed in linkers in the rest of the population.

The main significance of our analysis probably lies not in accounting for nonrandom locations of nucleosomes, but rather in validating the idea of an essentially random deposition of nucleosomes on DNA. The probability distributions we have derived are analogous to those for a one-dimensional liquid. (The only difference is that the one-dimensional "DNA space" is quantized rather than continuous.) Much as a liquid can adopt any shape and fill a container, so an essentially random array of nucleosomes can adjust to any size of DNA domain and package any sequence. Such a packaging mechanism has the great flexibility required by the functional diversity of chromosomes.

ACKNOWLEDGEMENTS

With gratitude and affection, we dedicate this article to Arthur Kornberg. His devotion to the advancement of knowledge is a continuing source of inspiration. We thank Michael Levitt for critical and stimulating discussion. This research was supported by Grant GM-30387 from the National Institutes of Health.

REFERENCES

1. Wittig, B. and Wittig, S. (1979) Cell **18**, 1173-1183.
2. Wu, C. (1980) Nature **286**, 854-860.
3. Levy, A., and Noll, M. (1980) Nucl. Acids Res. **8**, 6059-6068.
4. Bryan, P.N., Hofstetter, H. and Birnstiel, M.L. (1981) Cell **27**, 459-466.

5. Samal, B., Worcel, A., Louis, C. and Schedl, P. (1981). Cell **23**, 401-409.
6. Bloom, K. S. and Carbon, J. (1982) Cell **29**, 305-317.
7. Zhang, X.Y., Fittler, F. and Horz, W. (1983) Nucleic Acids Res. **11**, 4287-4306.
8. Lohr, D. (1983) Nucleic Acids Res. **11**, 6755-6773.
9. Edwards, C.A. and Firtel, R.A. (1984) J. Mol. Biol. **180**, 73-90.
10. Palen, T.E. and Cech, T.R. (1984) Cell **36**, 933-942.
11. Udvardy, A. and Schedl, P. (1984) J. Mol. Biol. **172**, 385-404.
12. Gottschling, D.E. and Cech, T.R. (1984) Cell **38**, 501-510.
13. Profitt, J.H. (1985) Mol. Cell. Biol. **5**, 1522-1524.
14. Almer, A., Rudolph, H., Hinnen, A., and Hörz, W. (1986) EMBO J. **5**, 2689-2696.
15. Almer, A. and Hörz, W. (1986) EMBO J. 5, 2681-2867.
16. Budarf, M.L., and Blackburn, E.H. (1986) J. Biol. Chem. **261**, 363-369.
17. Benezra, R., Cantor, C.R. and Axel, R. (1986) Cell **44**, 697-704.
18. Cartwright, I.L. and Elgin, S.C.R. (1986) Mol. Cell. Biol. **6**, 779-791.
19. Thoma, F. (1986) J. Mol. Biol. **190**, 177-190.
20. Richard-Foy, H. and Hager, G.L. (1987) EMBO J. **6**, 2321-2328.
21. Fedor, M.J., Lue, N.F. and Kornberg, R.D. (1988) J. Mol. Biol., in press.
22. Szent-Györgyi, C., Finkelstein, D.B. and Garrard, W.T. (1987) J. Mol. Biol. **193**, 71-80.
23. Chao, M.V., Gralla, J. and Martinson, H.G. (1979) Biochemistry **18**, 1068-1074.
24. Simpson, R.T. and Stafford, D.W. (1983) Proc. Natl. Acad. Sci. USA **80**, 51-55.
25. Ramsay, N., Felsenfeld, G., Rushton, B.M. and McGhee, J.D. (1984) EMBO J. **3**, 2605-2611.
26. Drew, H.R. and Travers, A.A. (1985) J. Mol. Biol. **186**, 773-790.
27. Kornberg, R.D. (1977) Ann. Rev. Biochem. **46**, 931-954.
28. Kornberg, R.D. (1981) Nature **292**, 579-580.
29. Nedospasov, S.A. and Georgiev, G.P. (1980) Biochem. Biophys. Res. Commun. **92**, 532-539.
30. Elgin, S.C.R. (1981) Cell **27**, 413-415.
31. Elgin, S.C.R. (1984) Nature **309**, 213-214.
32. Costlow, N.A., Simon, J.A. and Lis, J.T. (1985) Nature **313**, 147-149.
33. Emerson, B.M. and Felsenfeld, G. (1984) Proc. Natl. Acad. Sci. USA **81**, 95-99.
34. Emerson, B.M., Lewis, C.D. and Felsenfeld, G. (1985) Cell **41**, 21-30.
35. Jongstra, J., Reudelhuber, T.L., Oudet, P., Benoist, C., Chae, C.B., Jeltsch, J.M., Mathis, D.J. and Chambon, P. (1984) Nature **307**, 708-714.
36. Thompson, T. and Fan, H. (1985) Mol. Cell. Biol. **5**, 601-609.
37. Picard, D. and Schaffner, W. (1984) Nature **307**, 80-82.
38. Pospelov, V.A., Klobeck, H.G. and Zachau, H.G. (1984) Nucleic Acids Res. **12**, 7007-7021.
39. Zaret, K.S. and Yamamoto, K.R. (1984) Cell **38**, 29-38.
40. Briggs, M.R., Kadonaga, J.T., Bell, S.P. and Tjian, R. (1986) Science **234**, 47-52.
41. Payvar, F., DeFranco, D., Firestone, G.L., Edgar, B., Wrange, O., Okret, S., Gustafsson, J-A. and Yamamoto, K.R. (1983) Cell **35**, 381-392.
42. Bram, R.J. and Kornberg, R.D. (1985) Proc. Natl. Acad. Sci. USA **82**, 43-47.
43. Bram, R.J. and Kornberg, R.D (1987) Mol. Cell Biol. **7**, 403-409.
44. Buchman, A.R., Kimmerly, W.J., Rine, J. and Kornberg, R.D. (1988) Mol. Cell. Biol. **8**, 210-225.
45. Bellard, M., Dretzen, G., Bellard, F., Oudet, P. and Chambon, P. (1982) EMBO J. **1**, 223-230.

46. Workman, J.L. and Roeder, R.G. (1987) Cell **51**, 613-622.
47. Lorch, Y., LaPointe, J.W. and Kornberg, R.D. (1987) Cell **49**, 203-210.
48. Thoma, F. (1988) Architecture of Eukaryotic Genes (G. Kahl, ed.), VCH, Federal Republic of Germany, pp. 269-280.