The pitch of chromatin DNA is reflected in its nucleotide sequence

(deformation of DNA/dinucleotides/correlation analysis)

EDWARD N. TRIFONOV AND JOEL L. SUSSMAN

Polymer Department and Department of Structural Chemistry, The Weizmann Institute of Science, Rehovot, Israel

Communicated by David R. Davies, March 28, 1980

ABSTRACT   A correlation analysis of chromatin DNA nucleotide sequences reveals the clear tendency of some of the dinucleotides to be repeated along the sequences with periods of 3 and about 10.5 bases. This latter period, which is equal within experimental error to recent estimates of the pitch of the DNA double helix [Wang, J. (1979) Proc. Natl. Acad. Sci. USA 76, 200–203; Trifonov, E. & Bettecken, T. (1979) Biochemistry 18, 454–456] is interpreted as a reflection of the deformational anisotropy of the DNA molecule that facilitates its smooth folding in chromatin.

DNA of eukaryotic cells compacts severely when it folds into chromosomes. The elementary structural unit of chromatin is the nucleosome (1, 2), which consists of a histone protein core enveloped by DNA (3). One of the possible ways discussed recently in the literature that DNA can fold is by smooth bending with its deformation uniformly distributed along the length of the molecule (4–6).

Because the sequence of base pairs along a natural DNA double helix varies, the molecule in some aspects is anisotropic. This anisotropy might result in a local preference of the DNA molecule to be bent in a specific direction. For example, if adjacent base pairs normally were slightly nonparallel, this would cause some bending of the DNA axis. Thus, these two nonparallel base pairs could serve as a kind of wedge, changing the direction of the DNA axis. The orientation of the wedge depends on its position along the double helix of DNA, which is the same after each full turn of the helix. Therefore, to amplify the bending of the molecule in the same direction, such wedges could be inserted at regular intervals—multiples of the pitch of the DNA double helix (Fig. 1). This could facilitate the unidirectional bending of DNA in chromatin without destroying the base-stacking interactions. If some of the 16 possible combinations of adjacent base pairs are not strictly parallel, then the corresponding dinucleotides might have the tendency to

be positioned at regular intervals—every 10 or so bases—along the chromatin DNA sequences.

A correlation analysis of available sequences of eukaryotic DNAs and mRNAs (7–26) as well as of some viral DNAs (27–35) that are known to be folded in chromatin-like structures (36, 37) shows that some of the dinucleotides indeed exhibit this periodicity, although it is well hidden. The period is estimated to be 10.5 ± 0.2 bases, which is consistent with recent measurements and estimations of the pitch of DNA in dilute solution and in chromatin (6, 38).

METHOD AND RESULTS

A straightforward way to find out if any element of a nucleotide sequence tends to be periodically repeated is to calculate the corresponding positional autocorrelation function. One examines the frequencies of occurrence of the same elements at different distances from each other along the nucleotide sequence. The periodicity, if present, will result in higher frequencies of occurrence at distances that are multiples of the period.

In a typical example of such an autocorrelation function (Fig. 2A), all the distances between dinucleotides T-G in the simian virus 40 (SV40) DNA sequence (27) are scored, up to 35 bases along the sequence. In the distribution, distances that are multiples of 3 bases are more frequent and no other periodicities seem to be present. However, one could argue that there might be some weak periodicity masked by variations in the occurrences caused by the limited statistical ensemble when analyzing only a single dinucleotide. If there were some universally preferred distances, then they might be revealed by summation of the autocorrelation functions for all 16 dinucleotides. The result of this summation for the SV40 DNA sequence is shown in Fig. 2B, in which the 3-base period is more obvious. There also is a weak periodicity of about 10 bases, expressed as modulation of the 3-base pattern (see maxima at 9, 21, and 30 bases). To be sure that the effect is not a result of random fluctuations
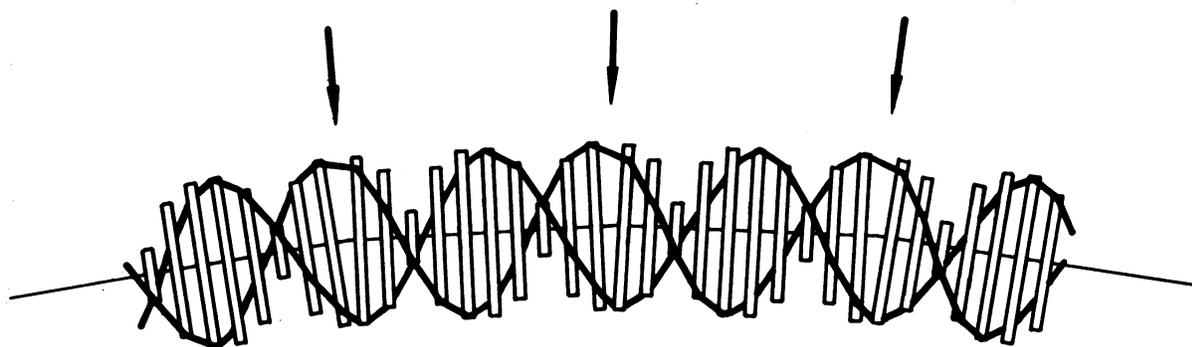
Abbreviation: SV40, simian virus 40.

FIG. 1.   Schematic illustration of the unidirectional bending of a DNA molecule by the regular insertion of a nonparallel set of adjacent base pairs (arrows).

3816

Biochemistry: Trifonov and Sussman

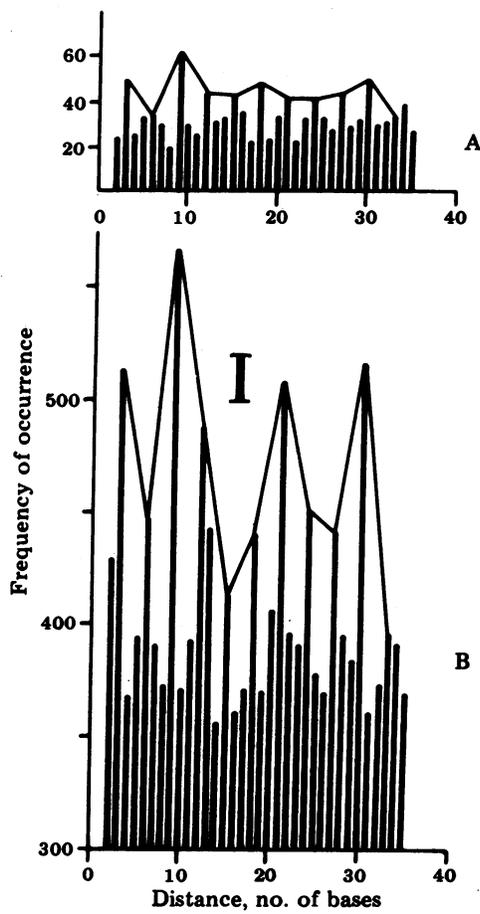*Proc. Natl. Acad. Sci. USA 77 (1980)* 3817

FIG. 2. Autocorrelation functions for dinucleotides of the SV40 DNA sequence. Points, corresponding to distances that are multiples of 3 bases, are connected by envelope curves. (*A*) Dinucleotide T-G; (*B*) sum of autocorrelation functions for all 16 dinucleotides. Heavy vertical bar corresponds to standard deviation estimated as the square root of the average frequency of occurrence. The *Y* axis origin is omitted for clarity.

Table 1. Chromatin nucleotide sequences used for the correlation analysis

Sequence	Length, bases	Ref.
Eukaryotes		
Ovalbumin gene	2368	7
Ovalbumin mRNA (portion not overlapping with above sequence)	1329	8
Histone (H2B and H3) genes of *S. purpuratus*	2033	9
Histone 2A gene of *P. miliaris*	734	10
Noncoding portion of histone 2A gene of *S. purpuratus* (portion not overlapping with above sequence)	233	9
Histone 4 gene of *P. miliaris*	561	10
Histone 1 gene of *P. miliaris*	258	10
Mouse Igκ chain mRNA (J cluster)	1736	11
Mouse Igκ light chain mRNA (constant and 3'-noncoding region)	532	12
Variable region of mouse IgλII light chain gene	726	13
Mouse IgλI light chain gene (clone 303)	702	14
Mouse IgλI light chain gene (clone 99; portion not overlapping with above sequence)	240	14
Mouse IgGγ1 heavy chain mRNA	458	15
Mouse β-globin major gene	1567	16
Rabbit β-globin mRNA	589	17
Rabbit α-globin mRNA	551	18
Bovine corticotropin/β-lipotropin precursor mRNA	1083	19
Iso-1-cytochrome *c* gene of *Saccharomyces cerevisiae*	857	20
Rat growth hormone mRNA	775	21
Human growth hormone mRNA	769	22
Human chorionic somatomammotropin mRNA	507	23
Yeast rDNA spacer	380	24
Rat preproinsulin mRNA	300	25
5S DNA of *Xenopus laevis*	273	26
Viruses		
SV40 genome	5226	27
Polyoma virus genome, early region	3013	28
Polyoma virus genome, late region	2370	29
Adenovirus 5 genome, transforming fragment	2810	30, 31
Adenovirus 5 genome, right-hand terminus	1078	32
Adenovirus 2, *Eco*RI F fragment of DNA	1743	33
Adenovirus 2, fiber mRNA	448	34
Adenovirus 2, hexon mRNA	240	35

of the frequencies, we calculated similar autocorrelation functions for 32 known contiguous pieces of nucleotide sequences, with a total length of about 36,000 bases (7–35), and summed the functions to enhance any regular component, if present. The result of the summation is shown in Fig. 3*A*. The sequences used for the calculation are listed in Table 1 (and can be provided in computer-readable form, if requested). We considered all these DNA sequences as chromatin-bound, although some of them—e.g., the region near the origin of replication in SV40 DNA—might not be involved in the regular chromatin structure (ref. 39; G. Kaufmann, personal communication). To avoid duplication, only nonoverlapping portions of related sequences were used. Some of the sequences published by different groups are nearly identical (e.g., histone genes, refs. 9 and 10). In such cases our choice was somewhat arbitrary, with preference to longer sequences. All the sequences analyzed corresponded to the coding strands except for the SV40 DNA sequence, in which the noncoding strand of the late region was used for the sake of physical continuity with the coding strand of the early region. The choice of the strand, however, does not matter, because the periodicity present in one strand is complementarily reflected in the other one.

The pattern in Fig. 3*A* looks essentially the same as that in Fig. 2*B*, both having a clear periodicity of about 10 bases. The periodicity can be improved in terms of relative amplitude of the variation if one considers that not all of the 16 dinucleotides

are contributing equally to the period. We estimated which of them display the strongest variations of occurrences at multiples of 3-base distances, as compared with expected variations for random sequences. Dinucleotides G-G, T-A, T-G, and T-T were found to be the "strongest" contributors to the variations. Summing the autocorrelation functions for only these dinucleotides for 32 sequences, we obtained a distribution (Fig. 4) in which the distances scored extend up to 100 bases and some additional small peaks are seen—for example, at about 54, 72, 84, and 93–96 bases. The cosine wave, which fits best to the envelope curve in Fig. 4, has a period of 10.5 ± 0.2 bases.

DISCUSSION

This analysis demonstrates that at least some of the 16 dinucleotides tend to be periodically distributed along the DNA that is associated with chromatin. The period, about 10.5 bases, appears to be equal, within experimental error, to the pitch of DNA in chromatin—10.33 to 10.40 base pairs (6)—estimated from data on the variation of the sensitivity to nuclease digestion
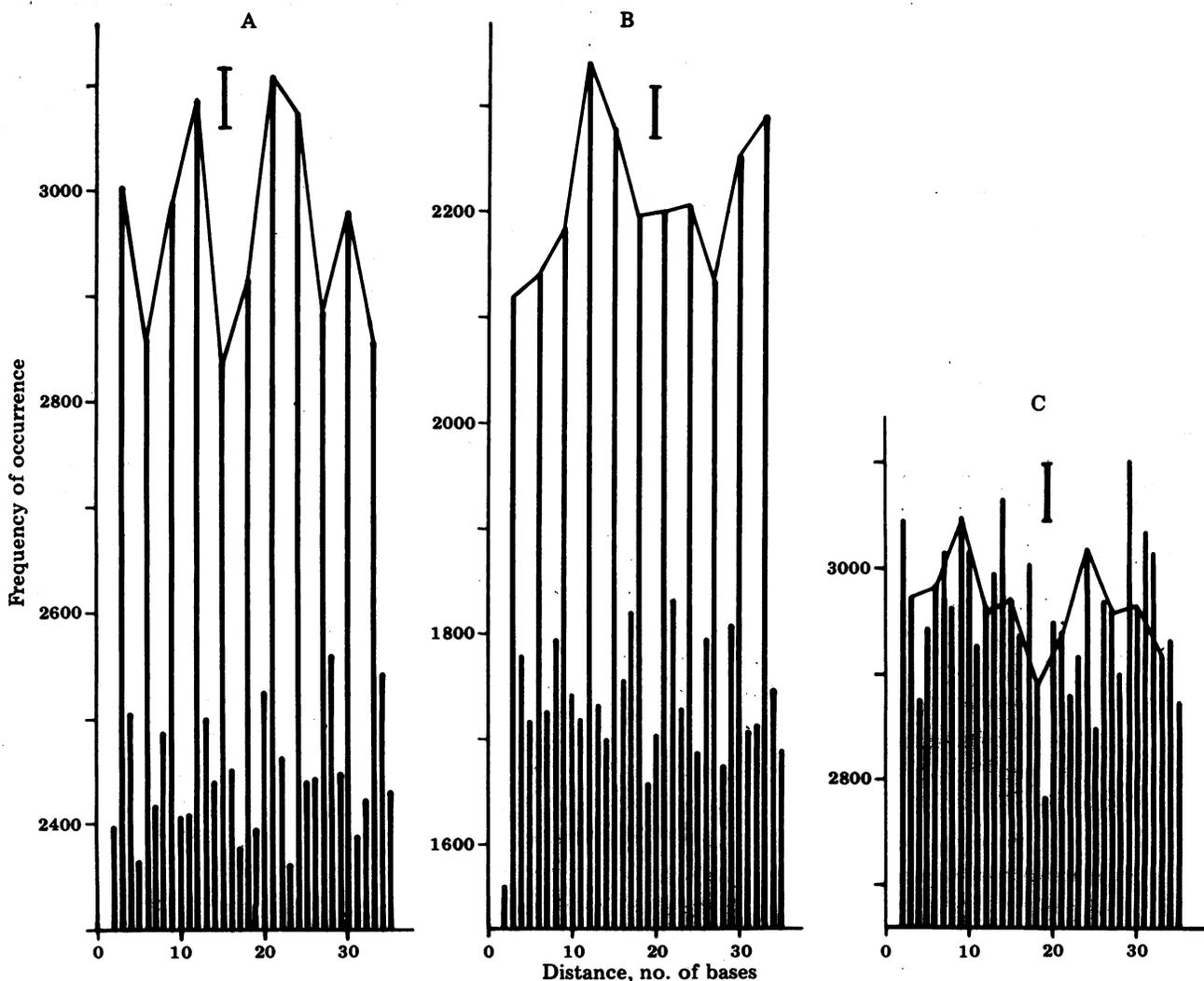
FIG. 3.   Sum of the autocorrelation functions for all 16 dinucleotides, calculated for chromatin sequences (A), nonchromatin prokaryotic sequences (B), and a random sequence (C). Envelope curves (as in Fig. 2) pass through the points that correspond to multiples of three-base distances. The Y axis origins are omitted for clarity. Because the autocorrelation functions for dinucleotides of the chromatin sequences have a slowly decaying component (e.g., as in Fig. 2B), this component was subtracted to make the oscillating part clearer. This was done by subtracting the deviation of each ordinate of the original distribution from the "running average" of seven adjacent points. The subtraction also was made for the cases of nonchromatin and random sequences for purposes of uniformity. The vertical bar in each graph corresponds to estimated standard deviation. The chromatin sequences used for the analysis are listed in Table 1; their total length is 36,000 bases. The nonchromatin prokaryotic sequences are listed in the text; their total length is 30,000 bases. The computer-generated random sequence has the same dinucleotide composition as SV40 DNA; the total length is 42,000 bases.

along the nucleosomal DNA (40, 41). This coincides numerically with experimentally measured DNase I intercleavage distance for DNA in chromatin—10.3 to 10.4 bases (42, 43). Thus, as one could predict from *a priori* DNA anisotropy considerations, the pitch of chromatin DNA is reflected in its nucleotide sequence. The value obtained, 10.5 ± 0.2 bases, can be considered as another independent estimate of the pitch of DNA in chromatin, although less accurate. Interestingly, the pitch of DNA in dilute solution appears to be the same—10.4 ± 0.1 base pairs (38).

The same analysis was applied as well to prokaryotic (nonchromatin) nucleotide sequences: full or partial nucleotide sequences of bacteriophages MS2 (44–46), φX174 (47), G4 (48), FD (49), and λ (50, 51), of *Escherichia coli* (52–54) and of plasmid pBR322 (55), with a total length of about 30,000 bases. The result of summing the corresponding autocorrelation functions is shown in Fig. 3B. The prokaryotic sequences exhibit a strong 3-base periodicity of about the same relative amplitude as for the chromatin sequences in Fig. 3A, but the 10.5-base

periodicity seems not to be present. A few of the prokaryotic sequences examined separately show some periodicity (of about 10 bases) which is not convincing statistically.

The variations in the random sequence of the same dinucleotide composition as in SV40 DNA are completely random (Fig. 3C).

The 10.5-base periodicity found in the case of SV40 DNA seems to be more pronounced than the average for all the chromatin sequences (e.g., compare relative amplitudes of the "signal" in Figs. 2B and 3A). This could mean that some portions of the sequences are not involved in tight folding in the chromatin.

The amount of dinucleotides participating in the formation of the 10.5-base periodical variation can be estimated from the amplitude, about 3%, of the autocorrelation function of the actual oscillation (Fig. 3A). The amplitude of the actual signal is equal to the square root of this value—i.e., about 20%. Therefore, approximately every fifth dinucleotide on the average is contributing to the 10.5-base periodicity.

Biochemistry: Trifonov and Sussman

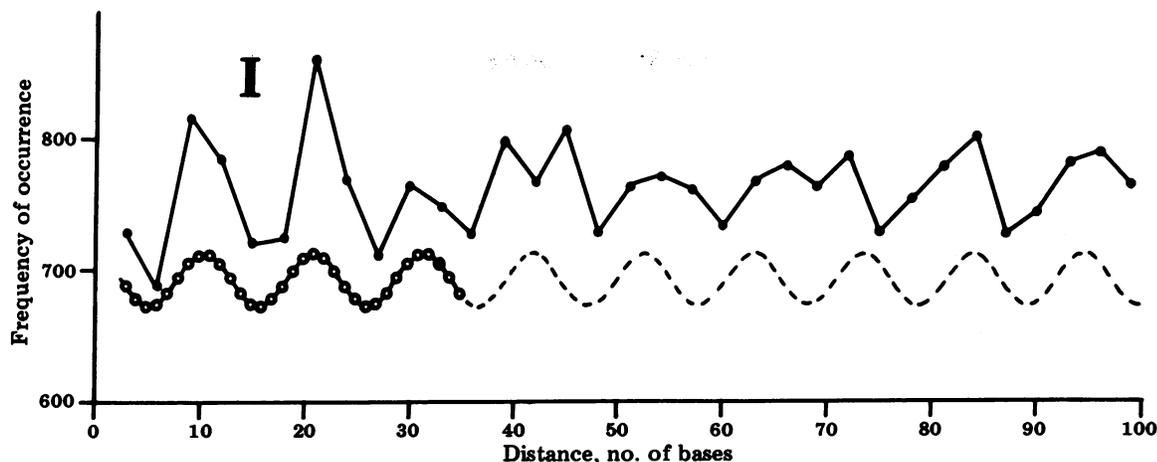*Proc. Natl. Acad. Sci. USA 77 (1980)*   3819

FIG. 4. Sum of the autocorrelation functions for the "strongest" dinucleotides G-G, T-A, T-G, and T-T. The "strength" of each dinucleotide was estimated as the mean dispersion for the corresponding frequencies of its occurrence (at distances of multiples of 3 bases), related to the square root of its average frequency. One strand of the DNA molecule with pitch 10.5 base pairs is shown schematically for comparison. Y axis origin is omitted.

If one does a similar correlation analysis of chromatin DNA, using mononucleotides instead of dinucleotides, their autocorrelation functions display as well the weak periodicity of about 10 bases, but the relative amplitude of the variation is about half of that for dinucleotides. This weak periodicity of the mononucleotides seems to be a reflection of the stronger variations for dinucleotides.

The 10.5-base oscillation found for dinucleotides of chromatin DNA reflects just the tendency of some dinucleotides to be repeated with this period along the sequences rather than a perfect repeating pattern, which could seriously disturb the genetic message. A way to lessen the interference is to use only each degenerate (third) nucleotide of the triplet coding frame to form the most acceptable dinucleotides in terms of the preferred unidirectional deformation of the DNA molecules. Probably, this is the reason why the 10.5-base periodicity is expressed as the modulation of occurrences at multiples of 3-base distances. The 3-base oscillation seen in the autocorrelation functions for dinucleotides could be related to the frequently observed preferences of some mononucleotides to be in the third positions of the triplets (18, 27). Another contribution to the 3-base periodicity could be due to some amino acids being used more frequently in proteins coded by the sequences analyzed. If this were the case, then the first and second positions of the codons should contribute to the 3-base periodicity also. (We are grateful to one of the referees for this alternative interpretation.)

Do noncoding regions contribute to the 10.5-base periodicity as well as coding ones? This intriguing question can be answered by analyzing these two subsets of chromatin sequences separately. Although the signal-to-noise ratio for each subset is lower than for the total ensemble, preliminary results of the analysis indicate that both coding and noncoding chromatin sequences contribute to the 10.5-base periodicity.

The correlations found decay with distance (Fig. 2B). This could mean that the 3- and 10.5-base periodicities are interrupted by phase shifts occurring once per 50–70 bases, possibly caused by erroneous deletions or insertions during sequencing procedures (16), or by splicings for mRNA and cDNA sequences.

If the 10.5-base periodicity found is a reflection of the specific deformational anisotropy of chromatin DNA, then further analysis of intercorrelations between dinucleotides might reveal which portions of the DNA sequences are facing into the histone core and which are pointing outside.

1. Kornberg, R. D. (1977) *Annu. Rev. Biochem.* **46**, 931–954.
2. Felsenfeld, G. (1978) *Nature (London)* **271**, 115–122.
3. Pardon, J. F., Worcester, D. L., Wooley, J. C., Tatchell, K., Van Holde, K. E. & Richards, B. M. (1975) *Nucleic Acids Res.* **2**, 2163–2176.
4. Sussman, J. L. & Trifonov, E. N. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 103–107.
5. Levitt, M. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 640–644.
6. Trifonov, E. N. & Bettecken, T. (1979) *Biochemistry* **18**, 454–456.
7. Robertson, M. A., Staden, R., Tanaka, Y., Catterall, J. F., O'Malley, B. W. & Brownlee, G. G. (1979) *Nature (London)* **278**, 370–372.
8. McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Givol, D., Fields, S., Robertson, M. & Brownlee, G. G. (1978) *Nature (London)* **273**, 723–728.
9. Sures, I., Lowry, J. & Kedes, L. H. (1978) *Cell* **15**, 1033–1044.
10. Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H. O. & Birnstiel, M. L. (1978) *Cell* **14**, 655–671.
11. Sakano, H., Hüppi, K., Heinrich, G. & Tonegawa, S. (1979) *Nature (London)* **280**, 288–294.
12. Hamlyn, P. H., Brownlee, G. G., Cheng, C.-C., Gait, M. J. & Milstein, C. (1978) *Cell* **15**, 1067–1075.
13. Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1485–1489.
14. Bernard, O., Hozumi, N. & Tonegawa, S. (1978) *Cell* **15**, 1133–1144.
15. Rogers, J., Clarke, P. & Salser, W. (1979) *Nucleic Acids Res.* **6**, 3305–3321.
16. Konkel, D. A., Tilghman, S. M. & Leder, P. (1978) *Cell* **15**, 1125–1132.
17. Efstratiadis, A., Kafatos, F. C. & Maniatis, T. (1977) *Cell* **10**, 571–585.
18. Heindell, H. C., Liu, A., Paddock, G. V., Studnicka, G. M. & Salser, W. A. (1978) *Cell* **15**, 43–54.
19. Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Cohen, S. N. & Numa, S. (1979) *Nature (London)* **278**, 423–427.
20. Smith, M., Leung, D. W., Gillam, S., Astell, C. R., Montgomery, D. L. & Hall, B. D. (1979) *Cell* **16**, 753–761.
21. Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D. & Goodman, H. M. (1977) *Nature (London)* **270**, 486–494.
22. Martial, J. A., Hallewell, R. A., Baxter, J. D. & Goodman, H. M. (1979) *Science* **205**, 602–607.
23. Shine, J., Seeburg, P. H., Martial, J. A., Baxter, J. D. & Goodman, H. M. (1977) *Nature (London)* **270**, 494–499.

24. Skryabin, K. G., Zakhar'ev, V. M. & Baev, A. A. (1978) *Dokl. Biochem.* **241**, 240–243.
25. Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W. J. & Goodman, H. M. (1977) *Science* **196**, 1313–1319.
26. Miller, J. R., Cartwright, E. M., Brownlee, G. G., Fedoroff, N. V. & Brown, D. D. (1978) *Cell* **13**, 717–725.
27. Reddy, V. B., Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. & Weissman, S. M. (1978) *Science* **200**, 494–502.
28. Friedmann, T., Esty, A., LaPorte, P. & Deininger, P. (1979) *Cell* **17**, 715–724.
29. Soeda, E., Arrand, J. R., Smolar, N., Walsh, J. E. & Griffin, B. E. (1980) *Nature (London)* **283**, 445–453.
30. Van Ormondt, H., Maat, J., De Waard, A. & Van der Eb, A. J. (1978) *Gene* **4**, 309–328.
31. Maat, J. & Van Ormondt, H. (1979) *Gene* **6**, 75–90.
32. Steenbergh, P. & Sussenbach, J. S. (1979) *Gene* **6**, 307–318.
33. Galibert, F., Herisse, J. & Courtois, G. (1979) *Gene* **6**, 1–22.
34. Zain, S., Sambrook, J., Roberts, R. J., Keller, W., Fried, M. & Dunn, A. R. (1979) *Cell* **16**, 851–861.
35. Akusjärvi, G. & Pettersson, U. (1979) *Cell* **16**, 841–850.
36. Germond, J. E., Hirt, B., Oudet, P., Gross-Bellard, M. & Chambon, P. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 1843–1847.
37. Sergeant, A., Tigges, M. A. & Raskas, H. J. (1979) *J. Virol.* **29**, 888–898.
38. Wang, J. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 200–203.
39. Varshavsky, A., Sundin, O. & Bohn, M. (1979) *Cell* **16**, 453–466.
40. Simpson, R. & Whitlock, J. (1976) *Cell* **9**, 347–353.
41. Noll, M. (1977) *J. Mol. Biol.* **116**, 49–71.
42. Prunell, A., Kornberg, R. D., Lutter, L., Klug, A., Levitt, M. & Crick, F. H. C. (1979) *Science* **204**, 855–858.
43. Lutter, L. (1979) *Nucleic Acids Res.* **6**, 41–56.
44. Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. (1972) *Nature (London)* **237**, 82–88.
45. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Merregaert, J., Min Jou, W., Raeymaekers, A., Volckaert, G., Ysebaert, M., Van de Kerckhove, J., Nolf, F. & Van Montagu, M. (1975) *Nature (London)* **256**, 273–278.
46. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. & Ysebaert, M. (1976) *Nature (London)* **260**, 500–507.
47. Sanger, F., Coulson, A. R., Friedmann, R., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A., III, Slocombe, P. M. & Smith, M. (1978) *J. Mol. Biol.* **125**, 225–246.
48. Godson, G. N., Barrell, B. G., Staden, R. & Fiddes, J. C. (1978) *Nature (London)* **276**, 236–247.
49. Beck, E., Sommer, R., Auerswald, E. A., Kurz, C., Zink, B., Osterburg, G. & Schaller, H. (1978) *Nucleic Acids Res.* **5**, 4495–4503.
50. Schwarz, E., Scherer, G., Hobom, G. & Kössel, H. (1978) *Nature (London)* **272**, 410–414.
51. Scherer, G. (1978) *Nucleic Acids Res.* **5**, 3141–3156.
52. Brosius, J., Palmer, M. L., Kennedy, P. J. & Noller, H. F. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4801–4805.
53. Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. & Dennis, P. P. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1697–1701.
54. Sugimoto, K., Oka, A., Sugisaki, H., Takanami, M., Nishimura, A., Yasuda, Y. & Hirota, Y. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 575–579.
55. Sutcliffe, J. G. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 3737–3741.