Craig, R. K., Castello, P. A., & Keir, H. M. (1975) *Biochem. J. 145*, 233–240.
Crute, J. J., Wahl, A. F., & Bambara, R. A. (1986) *Biochemistry 25*, 26–36.
Foster, K. A., & Collins, J. M. (1985) *J. Biol. Chem. 260*, 4229–4235.
Fry, M., & Loeb, L. A. (1986) in *Animal Cell DNA Polymerases*, CRC Press, Boca Raton, FL.
Harwell, L., Kappler, J. W., & Marrack, P. (1976) *J. Immunol. 116*, 1379–1384.
Keng, P. C., Li, C. K. N., & Wheeler, K. T. (1981) *Cell Biophys. 3*, 41–56.
Khan, N. N., & Brown, N. C. (1985) *Mol. Cell. Biochem. 68*, 169–179.
Khan, N. N., Wright, G. E., Dudycz, L. W., & Brown, N. C. (1984) *Nucleic Acids Res. 12*, 3695–3706.
Khan, N. N., Wright, G. E., Dudycz, L. W., & Brown, N. C. (1985) *Nucleic Acids Res. 13*, 6331–6342.
Kornberg, A. (1980) *DNA Synthesis*, W. H. Freeman, San Francisco.
Krauss, S. W., & Linn, S. (1986) *J. Cell. Physiol. 126*, 99–106.
Lee, M. Y. W. T., Tan, C.-K., So, A., & Downey, K. M. (1980) *Biochemistry 19*, 2096–2101.
Lee, M. Y. W. T., Tan, C.-K., Downey, K. M., & So, A. G. (1981) *Prog. Nucleic Acid Res. Mol. Biol. 26*, 83–96.
Lee, M. Y. W. T., Toomey, N. L., & Wright, G. E. (1985) *Nucleic Acids Res. 13*, 8623–8630.
Loeb, L. A., & Kunkel, T. A. (1982) *Annu. Rev. Biochem. 51*, 429–457.
Luk, C. K., Keng, P. C., & Sutherland, R. M. (1985) *Cancer Res. 45*, 1020–1025.
Miller, M. A., Wang, T. S.-F., & Korn, D. (1987) *Biochemistry* (in press).
Mishell, B. S., & Shiigi, S. M. (1980) *Selected Method in Immunology 368*, W. H. Freeman, San Francisco.
Pedrali-Noy, G., Spadari, S., Miller-Faures, A., Miller, A. D. A., Kruppa, J., & Koch, G. (1980) *Nucleic Acids Res. 8*, 377–387.
Penefsky, H. S. (1971) *J. Biol. Chem. 252*, 2891–2899.
Prelich, G., Tan, C.-K., Kostura, M., Mathews, M. B., So, A. G., Downey, K. M., & Stillman, B. (1982) *Nature (London) 326*, 517–520.
Skarnes, W., Bonin, P., & Baril, E. (1986) *J. Biol. Chem. 261*, 6629–6636.
Spadari, S., & Weissbach, A. (1974) *J. Mol. Biol. 86*, 11–20.
Spanos, A., Sedwick, S. G., Yarranton, S. T., Hubscher, U., & Banks, G. R. (1981) *Nucleic Acids Res. 9*, 1825–1839.
Tan, C.-K., Castillo, C., So, A. G., & Downey, K. M. (1986) *J. Biol. Chem. 261*, 12310–12316.
Tan, E. M. (1982) *Adv. Immunol. 33*, 167–240.
Tanaka, S., Hu, S.-Z., Wang, T. S.-F., & Korn, D. (1982) *J. Biol. Chem. 257*, 8386–8390.
Thommes, P., Reiter, T., & Knippers, R. (1986) *Biochemistry 25*, 1308–1314.
Wahl, A. F., Crute, J. J., Sabatino, R. D., Bodner, J. B., Marraccino, R. L., Harwell, L. W., Lord, E. M., & Bambara, R. A. (1986) *Biochemistry 25*, 7821–7827.
Waqar, M. A., Evans, M. J., & Huberman, J. A. (1978) *Nucleic Acids Res. 6*, 1933–1946.

The DNA Sequence of the Human β-Globin Region Is Strongly Biased in Favor of Long Strings of Contiguous Purine or Pyrimidine Residues†

Michael J. Behe
Department of Chemistry, Lehigh University, Bethlehem, Pennsylvania 18015
Received March 19, 1987; Revised Manuscript Received July 9, 1987

ABSTRACT: The DNA sequence of the human β-globin region, comprising over 67 kilobase pairs, has been analyzed for the occurrence of strings of contiguous purine or pyrimidine residues. Tracts of 10 or more contiguous residues are found 4 times more frequently than would be expected with a random distribution of bases, so that a long string occurs at an average of every 250 base pairs. A survey of six other human gene sequences, totaling 86 kilobase pairs, shows a remarkably similar result. No such overrepresentation of contiguous purine or pyrimidine residues is found in the bacteriophages λ or T7.

It has been known for quite some time that double-stranded synthetic polydeoxynucleotides in which one strand is exclusively purine residues have conformations that are different from DNA in which purines and pyrimidines occur on both strands. The synthetic polymers poly(dA)·poly(dT), poly(dI)·poly(dC), and poly[d(A-I)]·poly[d(T-C)] have been shown to differ from heterogeneous sequence DNA in their X-ray fiber diffraction patterns (Leslie et al., 1980). Poly(dA)·poly(dT) is known to have a helical repeat that is different from bulk DNA (Peck & Wang, 1981; Rhodes & Klug, 1981), and poly(dG)·poly(dC) is 20-fold less flexible than heterogeneous sequence DNA (Hogan et al., 1983). Neither poly(dA)·poly(dT) nor poly(dG)·poly(dC) is able to form nucleosomal structures when challenged by histones (McGhee & Felsenfeld, 1980), and a short cloned region of $(dA{\cdot}dT)_{20}$ was seen to be excluded from the central portion of nucleosomes (Kunkel & Martinson, 1981). The smallest number of contiguous purine residues that is needed for a segment of DNA in a longer strand to acquire a polypurine-like conformation is currently not known but is probably on the order of 10 base pairs or fewer.

It has been speculated that DNA sequences that have the ability to occur in conformations different from the B form,

† This work was supported by Grant GM36343 from the National Institutes of Health and Grant DMB-8510719 from the National Science Foundation. M.J.B. is a recipient of Research Career Development Award CA01159 from the National Institutes of Health.

0006-2960/87/0426-7870$01.50/0

Table I: Genbank Files Used in the Search

Genbank file designation	size of file (bases)	description of file
HUMHBB	67256	human β-globin region on chromosome 11
HUMFIXG	38059	human factor IX gene
HUMHBA4	12847	human α-globin region on chromosome 14
HUMNGFB	11594	human β-nerve growth factor gene
HUMTBB5	8874	human β-tubulin gene
HUMPOMC	8658	human proopiomelanocortin gene
HUMRASH	6453	human C-HA-rasl protooncogene
MUSGKAL1	9433	mouse glandular kallikrein genes
MUSMHAB3	10000	mouse MHC class II H2-IA-β gene
RABIGKCA	5235	rabbit Ig κ gene
RABUG	3709	rabbit uteroglobin gene
CHKOVAL	9206	chicken ovalbumin gene
CHKY	8372	chicken Y protein gene
XENHBBI	2972	*X. laevis* larval β-1-globin
XENHBBC	1989	*X. laevis* major β-globin gene
DROGART	9623	*D. melanogaster* Gart gene
DROHSP7D1	5066	*D. melanogaster* heat-shock locus 87C1
LAMBDA	48502	bacteriophage λ
T7	39936	bacteriophage T7

such as Z DNA or cruciforms, may be used by cells in some biological functions (Wang et al., 1979; Lilley, 1980). It has always been implicitly assumed, however, that such sequences would be present as a very small percentage of total genomic DNA. In this paper, however, I report the results of a search of human genomic DNA sequences for occurrences of contiguous purines or pyrimidines ranging from 1 isolated purine (flanked by 2 pyrimidines) to occurrences of greater than 15 in a row. The results show that, for the β-globin region and for 6 other human genes that were surveyed, occurrences of strings of contiguous purine or pyrimidine residues equal to or exceeding 10 bases in length are found at a frequency 4–6 times greater than expected. Since this corresponds to a total of ~5% of the DNA that was examined, with an average of about one string per 170–250 base pairs, these elements may play a basic role in influencing the structure of chromatin in vivo.

Materials and Methods

DNA sequences found in the Genbank genetic sequence data bank (Release 40.0, February 1986) were searched by using a microcomputer with commercial sequence analysis programs (International Biotechnologies, Inc.) written by Pustell and Kafatos (1982a,b, 1984). A description of the regions searched, the size of the regions, and the Genbank file designations are shown in Table I. These were selected primarily because they represent the largest files in the Genbank data base for the particular organisms and thus would be expected to give the most reliable statistical information.

Each region was searched for occurrences of the sequences -Py-$(Pu)_n$-Py- and -Pu-$(Py)_n$-Pu-, with n ranging from 1 to 15, where Py is pyrimidine and Pu is purine. All regions were then searched for occurrence of the sequences -$(Pu)_{16}$- and -$(Py)_{16}$-, which yields all strings of 16 or greater. This search strategy leaves no region of the reported sequence unexamined. For the 67 kilobase pair β-globin region, the sequence and start site were recorded for all strings of 10 residues or greater for subsequent analysis and are listed in Table II.

Expected frequencies of -Py-$(Pu)_n$-Py- for a random distribution of bases were calculated as follows: expected number of occurrences = number of bases in data set × (fraction of pyrimidines)^2 × (fraction of purines)^n. Expected frequencies of -Pu-$(Py)_n$-Pu- were calculated analogously.

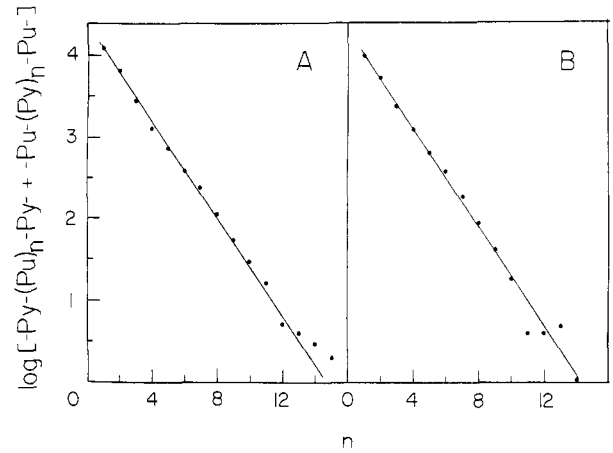

FIGURE 1: Logarithm of the sum of the number of -Py-$(Pu)_n$-Py- and -Pu-$(Py)_n$-Pu- sequences versus n, the number of contiguous purine or pyrimidine residues, for the DNA sequence of (A) bacteriophage λ and (B) bacteriophage T7.

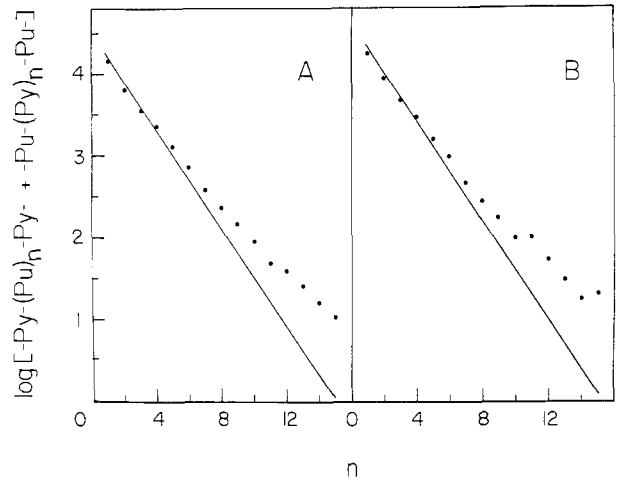

FIGURE 2: Logarithm of the sum of the number of -Py-$(Pu)_n$-Py- and -Pu-$(Py)_n$-Pu-sequences versus n, the number of contiguous purine or pyrimidine residues, for the DNA sequence of (A) the human β-globin region and (B) the sum of the six other human gene regions listed in Table I.

Results

The logarithm of the sum of the occurrences of the sequences -Py-$(Pu)_n$-Py- and -Pu-$(Py)_n$-Pu- is plotted versus n, the number of contiguous purine or pyrimidine residues, for the DNA sequences of bacteriophage λ and T7 in Figure 1A,B. The lines in the figures are not drawn to fit the data but are the predicted frequencies for the random occurrence of sequences of the types that were searched for. The data fit the predicted line very well, with a small amount of scatter for longer sequences with small expected frequencies of occurrence. A similar plot is shown in Figure 2A for the 67 kilobase pair DNA sequence of the human β-globin region. In this figure, however, the data do not fall along the predicted line but deviate in a positive direction as the length of contiguous purines and pyrimidines becomes greater. To determine if this was a feature peculiar to the β-globin region, six other human genes were analyzed, and the sum of the results is shown in Figure 2B. Again, the observed number of occurrences of long strings of contiguous purines or pyrimidines deviates in a positive direction from that expected for a random distribution of bases. The sum of all the sequences was used in order to provide a large sequence base. However, each individual gene shows the same behavior, albeit with a bit more scatter. Table III shows the expected and observed number of occurrences of strings of 10 or more contiguous purines and pyrimidines

Table II: Sequences of the 269 Strings Consisting of 10 or More Contiguous Purine or Pyrimidine Residues in the β-Globin Region[a]

SEQUENCE	START SITE
AAAAAAAAAAAAA	273
GGGAGGAGGG	851
AGAGGAAAAGGGG	1374
GAGAAGGAAA	1595
AAAAAGGAGGA	1724
GGAGAAGGGG	2968
AAGAGAGGAGAGGGAAGGA	4122
GAGAGGGAAA	5326
GAAAGAGAAG	5408
AAAAAAAAAAAAAGAAG	5903
GAGGGAGAAAAAGG	6404
GAGAAAGGAGAGGAGAAAGG	7700
AAAAAAAAGGAGAAG	8770
GGAAGAGGAAGG	9095
AGGAAAAGAAAA	9116
GAGAGGAAAGAAA	9314
GGGAGAAAAAA	9423
AGAGAGAAGA	9814
AAGAAAGGAAA	9927
AGAAAAAAAA	10045
GGGGAAAAAA	10059
GGAGGAAAAA	10456
AAAGAAAAAAA	13531
AAGAAGAAGAAAA	13628
AGAAAAAGAGG	13697
GGGGGAGGGAAAAA	13845
AGAGAGAGAGA	14730
GAGAGAAGGAGAGGA	14908
AGGAAAGAAAAG	15914
AAGGAGAGAAGAAAA	16214
AGAAAAAAAA	16751
GGGAGAAAAG	17640
AGAAGAGAGA	17745
AGGAGAAAGAAA	17917
AAAAAAAAAAAAAAAGAAAGAAAGAA--AAGAAAAAGAAAAAAGAAAAAA	18187
AGAAGGAGGAA	18255
AGAAAAAAAAAAGGA	18588
AAGAAGAGAG	18759
GGAAAGAGAGAGGA	18821
GGAAAAGGAGAA	19207
GAGGAGAAGG	19577
GAAAGAAGAA	20278
AGGGAGAGAGGGAA	20577
GGGGAAGGGGG	20611
AAAAGGAAAAAG	21142
GGGGAAAGAAGGAG	21360
AGAAGGGAGAG	21834
GAGAAGGAAGAAGGAGG	22166
GGGAGAGAGA	22493
AAGGAAGGGG	27083
AAGGAGGGGGAAG	29293
AGAGGGAAAG	29880
GAGGAAGGAGAA	32609
AAAAAAAAAAAAAGAGAGA	32708
AGAGAAAAAA	33304
AAAGAAAAAA	33352
AAGAAAAAAAA	33363
GGGAGAAGAAAA	34182
AAAGGGAAGAA	34200
GGAGGAGAAA	34622
AAGGGAGGGAAGGAAGGA	34663
GAGAAAAGAGAGG	35201
GGAAGAAAGA	36338
GAGGAAGGGG	36451
GAAGAAAAAG	36467
AAGGAGGAGA	36764
AGGAAGAGAG	37229
GAAGAAAGAGAAAAAAA	37340
GGGAGAAGGAAA	39118
AAAGGGAAGAA	39136
GGAGGAGAAA	39558
AAGGGAGGGAAGGAAGGA	39599
GAGAAAAGAGAGG	40137
AAAAGAGAGG	40573
GGGAAGAAAGA	41242
GAAGGGAAAG	41367
AAGGGAGAGA	41542
AGGAAGAGAA	42003
GAAGAAGGAGAAAAAAA	42118
AAGGAAGGAG	43170
AAGGGAAAGG	43540
GAGAGAAAAG	43712
AAAAAAAAAAAAAAAAAAAAAAAAG--AAGAAAGAAAGAAAAAAAAAG	45103
GGAGGGGAAGGAAG	45421
AGAGAGGAGAAG	45649
GAAAGAGGGA	45882
AAAAAGAGGA	46594
AAAAAAAAAA	46781
AAGAGAAGGA	46836
GGGGAGAAGAG	47245
AAGAAAAGAGAGAA	47457
AGGAAGGGGAAA	47504
AGAAAGGGAGG	47545
AAGAAAAAAAA	48555
GAAAAGAAGGAAA	48662
AAAGAAAGAGG	48813
GAAAAAAAAGA	49858
AAGGAAAGGA	50146
AGAAGGAGAAAA	50313
AGAGAGAGAGAGAGA	50491
GAGGGAAGGA	50624
AGAAGAGAAAAGAAAAAAAAGA	52313
AAAAGAAGGAGGAAG	54305
GAGGAGAAGA	54861
GGAAAGGAAAG	55727
GGGAAAAGAGAAAAG	56543
AGAAAGAAAA	57161
AAAGAGGAAG	57323
AAAAAAAAAGAG	57930
GGGAAGAGGGAGA	58289
GAAAGAGAGGGAGAGAGGAAGG--GAAGAGAGGA	58580
AGAAAGAAAAG	59820
AAGAGAAGGA	59927
AAAAAAGAAAA	59954
AGAAAAAGGAAAA	59971
AAAAAGAAAA	60450
GAAAGGAAGAAG	60834
AAGGGGAAAAAG	61076
GGGAGAAAGG	61099
AGAGAAAGAAGAG	61129
AGAGGAAAAAAA	61240
AAGGAGAAGA	61699
AGAGGGAGGG	61999
AGGAAGGGGAGAAG	62748
AAAAAGGGAA	63831
GAAGAAGGAAAAA	64268
AAAAGAAGAAAA	64957
AAAAGGAAAAA	65025
AAAAAAGAAAA	65076
AAAAAGAAGAAAA	65233
AAGAAAAAAG	65849
AAGGGAGAAA	66763
AAAAAAAAAAGAAAAAAAA--GGGGGGGGGGGG	67095
TCTCCCTCTC	10
TTCTTTTTCTT	1008
CCCTTTTCTCTCTCCC	1065
CTCTCCCTCCC	1220
TTTTTTTTTCCTTCTTTC	1352
TTTCTCTTTTT	1488
CTTCTTTCCTTT	1737
CCCCTTCTCT	5726
TTCCTCTTTCTC	6988
TCCTCTTTTCTT	7645
TTCTTTTCCCC	7974
CTTCCTTTTT	9845
TTTCTCTCTCT	10356
TTTTCCCTCCTT	10982
TTCTCCCCTTC	11806
TTTTTTTTTTTCTCC	12068
CTCCTTTTTTTTTTTTCT	12750
TCCTTCCCCT	13002
TTCCCCTCCCT	13047
TTTTTTTTTTTTTTTTTTTTT--TTTTTT	13076
TCCCTCCCTC	13439
TTTCTTTCCTTC	13737
CTCCTTTTTCCC	14607
CTCTTTTCCT	15331
CCTCCCCTTCCT	16997
TTTTTTTTTTTTTTTTTTTTT	17680
CTCCCTTCCC	18717
TTTCCTTTCCCT	19085
CCCTCCTTCT	21020
TCCTTCTTTCC	22881
TTTCCTTTCTTTT	23319
TTCCCTTTTCTCC	23463
TTTTCTCCTCT	23756
CTCTTCCTTTT	23792
TTCTTTTTTTT	24273
CTTTCTCTTT	24793
TTTTTTTTTTT	25281
TTCTCTCTTTCTCT	25358
CTTCTCTCCTCTTTTCTT	25562
TTCTCCTCTT	25687
TTCCCTTTCTTTT	25707
CTTTCCTTTT	25827
TCCCTCTCCT	26043
TTCCCCTCCCT	26090
TTCTTTCTCC	27156
CTTCCTCTCTTCCT	27861
TTTCTTTCTCTT	27891
TTTCTCTTCTTTCTC	29350
TTCTTTTTTCTTT	29442
TTTCTCCCCTC	29631
TCCCCCTTTTCCT	29947
TCTCTCTTCT	30002
TTTTCTTTTTCCTCC	30783
TCCCCCTTTTCC	31026
TCTTTTTTTCTCTCTCTCTTT--TTCTTCTCCCCC	31292
CTTTTTCCTT	31914
TCTTTTTCTTTCCTTTCTT	33016
CCCCTTCCTC	34294
CCTTTCCCCT	35268
TTCTTTTTTTT	36105
CCCCCTTCCCT	36422
TTTTTCTTTCC	36628
TTTCTCTTCCTTCT	37178
TTTTTTCTTC	37514
TTCTTTCTTT	37571
TTTCTTTCTTTCTTTTTTTTCCTC	37584
CTTCTTTCTCC	37648
CTTTCTTTCTTT	37741
TCCTTTTTTTTTT	37754
TTTTTTTCCTTTT	37954
CTTTTTCCTT	38912
CCCCTTCCCC	39230
CCTTTCCCCT	40204
CTCTTTTTTTT	41021
TTTCTCTTTCCTCCC	41951
CCCTTCTCTCT	42266
TTCTTTTTTCTTTCTTTCTTT--TTTCTCCTC	42363
CCCTTCCTTCTTTC	42415
CTTTCTTTCTTTTTT	42515
CCTTCTCTTTTT	42685
TTTTTTTTTCTT	42836
TCTTTTTCCC	42915
TCCTTTCCCC	43242
CTTCTTTTTCTTC	43956
TTCTTCTCTCT	44296
CTTCTCCCCC	44570
TTTTCCCCTT	44724
CTTTCTTCTTT	46159
TCTTTCCTTT	46488
TTTTCTTTTTT	47001
CTCCCTCCCC	47019
CTCTTTTTCT	47335
TTCTCTTTCT	48093
CCCTTTCTTCTCC	48841
CCTCTTTCTC	48873
TCTTTCCCTC	49582
TCTCTCTTTTCT	49889
TTCCTTCCTTCT	50010
CTTTTTTTTTTTTTTTTTCT	50948
CCTTTTCCTTTTT	51371
TTTCTCTTCC	51817
TTCCTCCCTC	52556
CTTTCTTCCCT	52903
TCTCTTCTCTT	53608
TCTCTTTTTTCCTCTTTTTTT--CCTTCCCTTCCCCTCTCTTC	53620
TTTCTTTTCTCTCCTCTTCTCT--TCTTTCCTCTCTTCCCTT--CCCTTTCTCTTTCTCTTC--CCTCTCTCCTCTCCCCTC--CTTTTTTCTCCTCTCCTCTCC	53661
TTTTTCCTTCTTCTCCTCC	53778
TCCTCTCTCTTCCCCTCTTCCTTCC--TTCCTTTCTCCCCCTCTTTCCCTC	53807
TCCTTCCTTTT	53867
TTCTTTCTTT	54088
CTTTTCTCTTTTT	55307
CTCTTCTCCC	55358
TTTTTTCCTTCCTC	55445
TCTTTCTCTCCC	55493
CCCCCTCCCTTC	55507
TTTTTCCTTCTCT	55520
CCTCTTCTCC	56170
TCTTTCTTCCTCCTTCCTTT	56787
CTTCCTCCTCTT	56867
TTTCCTTTCT	57358
TTTTTTTTCT	57384
TTCCTTTCTCC	57912
TTTTTCTTTCTC	58505
TCTCTTTTTT	59493
CCTTTTCCCCTCCT	61273
TTTTTTTTCTTTTCTT	61661
TTTTCTTTCCCCTTCTTTTCT	62708
CTTTCTTTTTTTTCTTCTCC	62884
CTTTCTTCTTTT	63196
TCTCTTTCTTTC	63250
TCTTCCTCCC	63521
CCTTTCCCCT	64166
CCCTTCTTCCCT	64310
TCCTCTCTCCC	64726
TTTTCCCCTTCCC	66162

[a] Starting bases for the genes of the region are the following: 19559 (ε gene); 34549 (γ_G gene); 39485 (γ_A gene); 54482 (δ gene); 62239 (β gene).

broken down for the individual gene regions.

To see if a bias in favor of long strings of purine or pyrimidine residues existed in the DNA of other eukaryotic species, 10 other genes were analyzed, 2 each from mouse, rabbit, chicken, *Xenopus laevis*, and *Drosophila melanogaster*, and the results are also shown in Table III. The higher organisms, mouse, rabbit, and chicken, show a bias of about the same magnitude as is found in human DNA. *Xenopus* and *Dro-*

Table III: Expected and Observed Number of Strings Consisting of 10 or More Contiguous Purine or Pyrimidine Residues

file name	expected no. of strings	obsd no. of strings	ratio (obsd/ expected)
HUMHBB	66	269	4.1
HUMFIXG	38	146	3.8
HUMHBA4	14	64	4.6
HUMNGFB	12	52	4.3
HUMTBB5	9	46	5.1
HUMPOMC	9	36	4.0
HUMRASH	7	44	6.3
MUSGKAL1	10	45	4.5
MUSMHAB3	10	41	4.1
RABIGKCA	6	24	4.0
RABUG	4	26	6.5
CHKOVAL	10	40	4.0
CHKY	9	46	5.1
XENHBBI	3	5	1.7
XENHBBC	2	5	2.5
DROGART	10	4	0.4
DROHSP7D1	5	14	2.8
LAMBDA	48	60	1.3
T7	40	33	0.8

Table IV: Sequence Composition of the 269 Strings Consisting of 10 or More Contiguous Purine or Pyrimidine Residues in the β-Globin Region[a]

	base composition	
purine strings	62% A	38% G
pyrimidine strings	62% T	38% C

	ratio of observed/expected dinucleotide frequencies							
purine	AA	AG	GA	GG	5′-A	5′-G	3′-A	3′-G
strings	1.01	0.98	0.98	1.01	0.98	1.04	0.97	1.04
pyrimidine	TT	CT	TC	CC	5′-T	5′-C	3′-T	3′-C
strings	1.03	0.96	0.97	1.04	1.00	1.00	0.96	1.07

[a] Number of strings containing 10 or more contiguous, identical bases: A + T, 12; G + C, 1.

sophila have a smaller frequency of occurrence of oligopurine strings. This may be due to problems in sampling, however. Reported sequences for lower organisms in general tend to be short and to consist mainly of coding sequences. Thus, the *Xenopus* files reported here may be too small to get an accurate indication of oligopurine string frequency. The DROGART file consists almost entirely of coding sequences, which may account for its low frequency of oligopurine strings. The *Drosophila* heat-shock sequence is less than 50% coding sequences and has a greater than expected occurrence of oligopurine strings.

An analysis of the sequences of the strings of the β-globin region is shown in Table IV. The 269 strings are 62% A + T and 38% G + C, compared to 61% A + T and 39% G + C for the β-globin region as a whole. The frequency of appearance of purine or pyrimidine dinucleotides in the tracts is very close to what is expected for random events. The 5′ and 3′ ends of the strings also show no preference for A or G in purine strings or T or C in pyrimidine strings. There are only 12 runs of contiguous A's or T's that are 10 residues in length or greater out of 269 strings. Besides 1 run of 12 G's in a longer run of 31 purines (start site = 67 095), there are no occurrences of stretches of more than 5 G's or C's in a row. There are only 3 strings of alternating -(AG)- or -(TC)- with a length of 10 residues or greater (1 of 10 base pairs and 2 of 14 base pairs). Thus, the bias in the regions examined appears to be for runs of purines or pyrimidines without regard to type or sequence.

In order to determine if there is any strong periodicity to the placement of the strings of the β-globin region along the DNA strand, the distance between nearest-neighbor strings was plotted in Figure 3 versus the frequency of occurrence. Again, the curve in the figure is not drawn to fit the data but is a calculated distribution for random placement of 269 elements on a line 67 256 units in length. The data fit the curve fairly closely, indicating that if there is a nonrandom location of strings it is not strongly periodic.

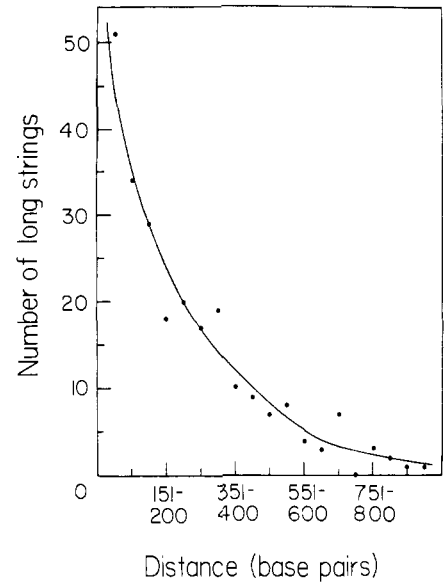

FIGURE 3: Number of strings of 10 or more contiguous purines or pyrimidines of the human β-globin region versus separation in base pairs from the nearest-neighbor string. The distance in base pairs between nearest neighbors was recorded for all 269 strings of the β-globin region. The number of strings with nearest neighbors falling in 50 base pair intervals (e.g., 0–50, 51–100, 101–150, etc.) were counted and plotted.

Table V: Site of Occurrence of Strings of the β-Globin Region as a Function of Length

length of string	site of occurrence			
	flanks (23 153 bp)	intergenic (36 901 bp)	introns (4 969 bp)	exons (2 205 bp)
10	24	57	7	4
11	15	32	2	0
12	17	18	4	0
13	7	15	4	0
14	3	10	3	0
15	4	7	0	0
>15	5	27	4	0

Table V shows the number of strings 10 residues or greater in the β-globin gene region broken down into flanking regions, intergenic region (the region contains 5 expressed genes: ε, γ_A, γ_G, δ, and β), coding regions, and introns. Most of the strings occur in the flanking and intergenic regions, but these are also the largest areas of the gene region. Only 4 strings occur in coding regions, and all of those were strings of 10, none longer. There are 8 occurrences of very long strings, 25 or greater in length, with no interruptions, another 3 that have interruptions of 1 base pair, and one "monster" region 170 pyrimidines in length (Poncz et al., 1980) with two single-base interruptions. Of these very long strings, 11 are in intergenic or flanking regions, 1 is in an intron (the second intron of the δ gene), and there are none in coding regions.

DISCUSSION

Characteristics of Oligopurine Strings. As seen in Figure 2A, the DNA sequence of the human β-globin region has a clear bias in favor of long strings of purines or pyrimidines. This bias is also found in the six other human genes that were

examined. Because Figure 2 uses a log scale, it is more difficult to appreciate that the bias increases as the length of the string increases, from a factor of about 3-fold greater than random for strings of 10 in a row to 10-fold for strings of 15, with an average of 4-fold greater than random for all strings of 10 or greater in length. This means that there is 1 string of 10 or greater about every 250 base pairs for a total of about 5% of the DNA examined.

The mouse, rabbit, and chicken genes that were examined (Table III) also had a strong sequence bias in favor of long strings of contiguous purine or pyrimidine residues, similar to human DNA. Although the *X. laevis* and *D. melanogaster* sequences showed a lower frequency of occurrence of oligopurine strings (Table III), it is difficult to determine if this is an accurate result or if it is due to the limited sequence information available for these organisms. It appears from the data presented here that the sequence bias may be a general phenomenon for higher eukaryotic organisms, but more data must be gathered before a conclusion can be reached concerning lower eukaryotes.

The strings of contiguous purine or pyrimidine residues found in the β-globin region have no apparent sequence preference. There is only a small number of strings with simple repetitive sequences longer than 10 base pairs: 12 strings of contiguous dA or dT residues, 1 of contiguous dG residues, and 3 of alternating -(AG)- or -(TC)- residues out of a total of 269 strings for the region. Table IV shows that the frequency of occurrence of purine or pyrimidine nearest neighbors is very close to that expected for random occurrence of either purine or pyrimidine. Furthermore, the frequency of occurrence of trinucleotide sequences in the strings is also very close to that expected for random mixing (data not shown). Therefore, I emphasize that the sequence bias in the human β-globin gene appears to be for strings of contiguous purine or pyrimidine residues *without regard to type or sequence.* This is an interesting result because it seems to preclude any simple explanation for the occurrence of the bias, such as unequal crossover between repetitive sequences or a large number of copies of one or several interspersed, highly repetitive DNAs.

In examining over 22 kilobases of the human β-globin region by electron microscopy of self-annealed globin clones, Coggins et al. (1980) found only five repetitive DNA sequences of average length 259 base pairs, two of which contained AluI restriction sites. The consensus sequence for the Alu family of dispersed repetitive DNA (Jelinek & Schmid, 1982) does not contain a 10 base pair string of contiguous purine or pyrimidine residues. The A-rich 3′-flanking sequence of the Alu family, usually of the general form $[N(A)_n]_m$, where N represents any nucleotide, n is usually less than 20, and m is usually less than 10, however, can contain such a string. I have searched the entire 67 kilobase pair β-globin region and found 12 sequences homologous to the Alu consensus sequence. Seven of the strings listed in Table II (start site = 5903, 18187, 32708, 45103, 50948, 58580, and 67095) occur at the 3′ end of Alu sequences. Additionally, another two strings (start site = 273 and 32609) occur within Alu sequences by deviation from the consensus sequence. Since these nine strings occur in a total of ~3600 base pairs of Alu DNA (12 × 300 base pairs), there is an average of one string per 400 base pairs of Alu DNA, less than the β-globin region as a whole.

Alu sequences account for about 40% of the short-period dispersed repeated human DNA sequences (Jelinek & Schmid, 1982). If other short-period repetitive sequences contributed a proportional number of oligopurine strings to the region, then a total of ~23 strings would be contained in the repetitive sequences. Thus, there are not enough repetitive sequences in the β-globin region to account for more than a small percentage of the 269 strings.

There is no readily apparent preferential placement of strings. Figure 3 shows that, to the limits of our ability to detect, the placement of strings fits closely to that expected for random occurrence. Long strings are seen to occur both in introns and in intergenic regions (Table V). No very long strings have been seen to occur in coding regions. In the β-globin region, only 4 strings of 10 occur in coding sequences. In 2 of these, the amino acid sequence (-Glu-Glu-Lys-Ala-) requires codons having 10 contiguous purines, and in the other 2 strings, purines are required in 8 of 10 nucleotide residues to give the observed amino acid sequence (-Gly-Gly-Glu-Thr-).

Possible Effects of the Observed Sequence Bias. It seems apparent from the work of Crothers' laboratory (Wu & Crothers, 1984; Koo et al., 1986) that sequences of four or more contiguous dA residues are in an altered conformation, as shown by their ability to cause bending of DNA fragments when the dA strings are phased every 10 base pairs. The ability of such strings to cause bending of DNA, however, is only a single effect of their occurrence in an altered conformation. The bias reported here in favor of long strings of purine or pyrimidine residues in the human β-globin region may be an exploitation of an altered structure of the strings to exert other effects.

What might be the effect of the bias in favor of long strings of purine or pyrimidine residues in the human β-globin region, and perhaps throughout many eukaryotic genomes? One possibility is that, since the bacteriophages λ and T7 do not show such a bias, the strings might be designed to affect nucleosome stability or placement. The frequency of occurrence of strings of purines 10 or more in length, about 1 every 250 base pairs for a total of ~5% of the DNA that was surveyed, is great enough to conceivably have a general effect on the chromatin structure of the entire region, not just on specialized elements such as promoter sites.

But how strong of an effect on nucleosome structure would oligopurine strings of ~10 base pairs have? There are examples of naturally occurring DNAs where nucleosome formation is known not to occur in vivo on a region containing oligopurine strings. A nuclease-sensitive region near the 5′ end of the chicken β-globin gene that was shown to be nucleosome free (McGhee et al., 1981) contains 4 oligopurine strings in a 200 base pair region, including a run of 18 contiguous guanosines, and a string of 32 purines with 1 interruption occurs in a region of the yeast TRP1ARS1 plasmid (nucleotides 10–43) that does not have a nucleosomal structure in vivo (Thoma et al., 1984). It is apparent, however, that nucleosomes can form over oligopurine strings, both in vitro and in vivo. A *Drosophila melanogaster* simple satellite DNA consisting of the repeating sequence -(AAGAG)- was seen to be packaged into nucleosomes in vivo (Levinger, 1985), and one clone of a number of chicken nucleosomal DNA fragments was seen to contain a long stretch of alternating -(AG)- (Satchwell et al., 1986). Additionally, we have successfully reconstituted the synthetic polypurine poly[d(AG)]·poly[d(TC)] using chicken histones by the direct mixing procedure (Jayasena and Behe, unpublished observations).

Satchwell et al. (1986) have analyzed the DNA sequences of 177 cloned nucleosomal fragments and reported that, while oligo(dA) strings did occur in all positions of nucleosomal DNA on various fragments, there was a strong preference of strings containing six or more contiguous adenosine residues

to be placed toward the ends of the fragment, avoiding the central dyad region. It seems likely that a similar result will be found for other long oligopurine strings: that although such strings can be packaged into nucleosomes if there is a sufficient driving force, they are less stable in a nucleosomal structure than other sequences.

Poljak and Gralla (1987) have recently performed competitive reconstitution of SV40 restriction fragments. SV40 contains one 15 and one 17 base pair oligopurine string, but the fragments containing the long strings reconstituted normally even though several other fragments did not. However, the choice of restriction enzymes leaves one string on the end of a 215 base pair fragment and the other string in the middle of a 610 base pair fragment so that it is possible for the strings to avoid being packaged into a nucleosome forming on the respective restriction fragments. To determine the effect of oligopurine strings on nucleosome formation, it will ultimately be necessary to quantitatively measure their affinity for nucleosome formation in competition with other DNA, and to determine the affinity as a function of the position of the oligopurine sequence along the nucleosome.

It must also be emphasized that, while oligopurine strings may eventually be shown to be important factors in the determination of nucleosome placement or stability, there are other factors that influence the positioning of nucleosomes. For example, several groups have demonstrated the precise alignment of a nucleosome in vitro on DNA sequences that have no long oligopurine strings (Simpson & Stafford, 1983; Ramsay et al., 1984; Nobile et al., 1986), and Blasquez et al. (1986) have shown the importance of protein factors in determining the average nucleosomal repeat length of the SV40 minichromosome.

Although an effect on nucleosome stability springs quickly to mind, the overabundant oligopurine strings of human DNA may have another role or roles. It is possible, for example, to imagine an effect of such elements on the higher order structure of chromatin or on the folding of chromatin into chromosomes during metaphase. Whatever the role of overabundant oligopurine strings in vivo, it is clear that in future work on the structure and function of eukaryotic DNA the ramifications of the sequence bias in favor of long strings of contiguous purine or pyrimidine residues will have to be kept in mind.

References

Blasquez, V., Stein, A., Ambrose, C., & Bina, M. (1986) *J. Mol. Biol. 191*, 97–106.

Coggins, L. W., Grindlay, G. J., Vass, J. K., Slater, A. A., Montague, P., Stinson, M. A., & Paul, J. (1980) *Nucleic Acids Res. 8*, 3319–3333.

Hogan, M., LeGrange, J., & Austin, B. (1983) *Nature (London) 304*, 752–754.

Jelinek, W. R., & Schmid, C. W. (1982) *Annu. Rev. Biochem. 51*, 813–844.

Koo, H. S., Wu, H. M., & Crothers, D. M. (1986) *Nature (London) 320*, 501–506.

Kunkel, G. R., & Martinson, H. G. (1981) *Nucleic Acids Res. 9*, 6869–6888.

Leslie, A. G. W., Arnott, S., Chandrasekaran, R., & Ratliff, R. L. (1980) *J. Mol. Biol. 143*, 49–72.

Levinger, L. (1985) *J. Biol. Chem. 260*, 11799–11804.

Lilley, D. M. J. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 6468–6472.

McGhee, J. D., & Felsenfeld, G. (1980) *Annu. Rev. Biochem. 49*, 1115–1156.

McGhee, J. D., Wood, W. I., Dolan, M., Engel, J. D., & Felsenfeld, G. (1981) *Cell (Cambridge, Mass.) 27*, 45–55.

Nobile, C., Nickol, J., & Martin, R. G. (1986) *Mol. Cell. Biol. 6*, 2916–2922.

Peck, L. J., & Wang, J. C. (1981) *Nature (London) 292*, 375–378.

Poljak, L. G., & Gralla, J. D. (1987) *Biochemistry 26*, 295–303.

Poncz, M., Schwartz, E., Ballantine, M., & Surrey, S. (1980) *J. Biol. Chem. 258*, 11599–11609.

Pustell, J., & Kafatos, F. C. (1982a) *Nucleic Acids Res. 10*, 51–59.

Pustell, J., & Kafatos, F. C. (1982b) *Nucleic Acids Res. 10*, 4765–4782.

Pustell, J., & Kafatos, F. C. (1984) *Nucleic Acids Res. 12*, 643–655.

Ramsay, N., Felsenfeld, G., Rushton, B. M., & McGhee, J. D. (1984) *EMBO J. 3*, 2605–2611.

Rhodes, D., & Klug, A. (1981) *Nature (London) 292*, 378–380.

Satchwell, S. C., Drew, H. R., & Travers, A. A. (1986) *J. Mol. Biol. 191*, 659–675.

Simpson, R. T., & Stafford, D. W. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 51–55.

Thoma, F., Bergman, L. W., & Simpson, R. T. (1984) *J. Mol. Biol. 177*, 715–733.

Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G., & Rich, A. (1979) *Nature (London) 282*, 680–686.

Wu, H. M., & Crothers, D. M. (1984) *Nature (London) 308*, 509–513.