Avoidance of base runs in switch regions of immune-system genes.

Pascale Perrin, Richard Grantham

Avoidance of Base Runs in Switch Regions of Immune-System Genes[1]

Pascale Perrin and Richard Grantham
Institut d'Evolution Moléculaire, Université Claude Bernard Lyon I

Mouse immunoglobulin (Ig) switch-region sequences are anti-"runny"; that is, they have a smaller amount of their total bases in homonucleotide tracts ("runs") than would be expected if each nucleotide in the sequence were a random selection from a pool of the composition of the region. The switch sequences involve the first intron of rearranged Ig heavy-chain genes; this intron differs strikingly from the succeeding ones, which are "runny" (have more bases than expected in runs). Switch regions are the only category of sequences so far found to be antirunny by statistical test. This sequence characteristic is related to the presence in switch sequences of repeating heteronucleotides. We suggest that the resulting base dispersion and increased complexity favor more specific interactions between sequences, which may be advantageous in recombinational processes such as switching and translocation.

Introduction

The genetics of the effector molecules of the immune system, antibodies and T-cell antigen receptors (Tcr), is characterized by unusual somatic recombination events. In brief, the production of an operational mRNA for either of the two polypeptides (light [L] and heavy [H]) making up an antibody or the components of the Tcr is preceded by a DNA rearrangement that juxtaposes the sequence coding for the variable (V) region to a J (joining) segment upstream of that coding for the constant (C) region. In the case of immunoglobulin-heavy (IgH) chains and of Tcr peptides (alpha, beta, or gamma) (Hood et al. 1985; Mori et al. 1985), another sequence, the diversity (D) segment, is also implicated in the rearrangement. These recombinations appear to involve repeated oligonucleotides that are separated by one or two turns of the DNA helix (Tonegawa 1983; Chien et al. 1984; Kavaler et al. 1984).

A second recombination event is seen in the IgH genes. The initial expression by the immature B cell is of mu chains, which are coded by the 5′ exons in the constant-heavy (CH) cluster. Some expression of delta chains, which are coded by the second exon in the cluster, is achieved by alternative splicing of a long pre–messenger RNA, in which the entire mu coding is eliminated with the first intron (Mather et al. 1984). On further maturation of the B cell, a "switch" may occur. This event involves the elimination of the DNA between the intron 5′ to the IgM gene (which codes for the mu chain) and the region 5′ to the newly expressed CH gene, including all intermediate CH-coding regions. The cell then ceases production of IgM and makes immunoglobulin of a different class (IgG, IgA, or IgE) but carrying the same V region

1. Key words: Ig switch sequences; runs; antirunniness; exons; introns; Ig classes A, E, G and M; recombination; translocation.

Address for correspondence and reprints: Richard Grantham, Institut d'Evolution Moléculaire, Université Claude Bernard Lyon I, 69622 Villeurbanne, France.

Mol. Biol. Evol. 5(2):141–153. 1988.

0737-4038/88/0502-0004$02.00

as that in the original IgM, thus conserving antigen specificity (Cushley and Williamson 1982).

In addition, chromosome translocations that appose a c-*onc* gene, usually c-*myc*, to regions coding Ig chains are often observed in lymphoid tumors. These translocations may occur with L-chain genes but are much more frequent in the H-chain cluster. They may be favored by either of the above recombination events (Klein and Klein 1985).

In this paper we describe the organization of certain structural sequence elements, notably "runs" in the untranslated regions 5′ to IgH-chain codons. These regions were compared with their analogues in other parts of the immune system, in particular with the major histocompatibility complex (MHC), and with appropriate zones of the c-*myc* gene. We find that, in contrast to nearly all other kinds of sequences studied, the regions 5′ to IgH genes involved in the switch mechanism have fewer homonucleotide tracts than would be expected on the basis of their base composition.

We have studied murine switch sequences (few switch sequences of other species are known) to try to answer the following question: How do these target zones for recombination differ from other sequences?

Gene sequences from three families of the immune system: immunoglobulins, T-cell receptors, and the MHC have been analyzed. Each Ig or Tcr chain contains a V and a C part. MHC genes (H-2 in mouse and HLA in man) code three classes of proteins—I, products of ubiquitous distribution on nearly all cells; II, products present on membranes of lymphoid cells; and III, components of complement—the first two of which are highly polymorphic surface glycoproteins (Müller-Eberhard 1975; Chaplin et al. 1983).

The term "run" is used herein to mean a homonucleotide sequence of at least two bases, contrary to the practice of David and Barton (1962), who call even single bases runs. Thus the dinucleotide AA is an adenine run of length 2, CCCC is a cytosine run of length 4, etc. For convenience, sequences containing more of such homonucleotides than expected on the basis of base content are called "runny" (Grantham et al. 1985). We show that runs are avoided in switch regions.

Nucleotide Sequences Studied

The largest possible sample has been taken from GenBank (release 46). Sequences missing in GenBank were entered directly into our information system ACNUC (Gouy et al. 1984; Grantham et al. 1985), which served for treatment of all the sequences. Description of the work sample appears in table 1. In this study we consider the entire intron preceding a CH-coding region as a switch zone.

Statistical Methods

Sequences have been differentiated according to content of runs for the following two reasons: (*a*) Previous work on switch regions revealed the existence of repeat structures based on a heteronucleotide repeating unit (Nikaido et al. 1981), suggesting the possible avoidance of runs. (*b*) A working hypothesis has been that runs may be a recognition element for recombination related to DNA conformation change, as suggested by work on homopurine/homopyrimidine tracts (Cantor and Efstradiatis 1984; Johnston and Rich 1985).

1. A statistical test for significance of the number of bases found in runs, given the sequence $\underline{A\ A}\ U\ G\ \underline{A\ A\ A}\ U\ G\ C\ U\ U\ \underline{A}$, was constructed as follows: according to the David and Barton (1962) test we use, the number of runs (NR) = 3 for base A

Table 1
Work Sample of Immune-System Sequences

Sequence Family, Species, and Region (Workfile)	No. of Sequences (No. of Nucleotides)				
	Exons	5′ NT End	3′ NT End	Intron	Switch Region
IG:					
Man:					
C (CONSTHUM)	8 (7.0 kb)	...	...	16 (3.1 kb)	
V (VARHUM)	9 (3.3 kb)	...	...	9 (1.0 kb)	
Mouse:					
C (CONSTMUR)	11 (9.0 kb)	...	...	15 (2.7 kb)	
(SWITMUS)					17 (9.7 kb)
Vg (VARGERM)	25 (8.6 kb)	11 (3.3 kb)	...	23 (2.7 kb)	
Vr (VARREC)	34 (12.9 kb)	...	...	14 (1.8 kb)	
Rabbit:					
C (ORIGC)	9 (5.2 kb)	...	...	...	
V (ORIGV)	4 (1.5 kb)	...	...	...	
Rat:					
C (RATIGC)	4 (2.5 kb)	...	...	...	
V (RATIGV)	1 (0.4 kb)	...	...	...	
Chicken:					
C (CHIGC)	2 (1.4 kb)	...	...	...	
V (CHIGV)	1 (0.3 kb)	...	...	...	
Caiman:					
V (CAIGV)	1 (0.5 kb)	...	...	...	
MHC:					
Man:					
HLA-I (HLAI)	3 (3.3 kb)	3 (0.9 kb)	3 (1.7 kb)	21 (5.5 kb)	
HLA-II (HLAII)	12 (8.9 kb)	3 (2.3 kb)	11 (4.9 kb)	10 (12.3 kb)	
HLA-III (HLAIII) ...	2 (3.3 kb)	...	...	10 (1.9 kb)	
Mouse:					
H2-I (H2)	6 (5.0 kb)	2 (0.5 kb)	3 (3.3 kb)	19 (7.8 kb)	
H2-II (H2II)	13 (9.7 kb)	5 (1.9 kb)	12 (7.2 kb)	19 (16.3 kb)	
H2-III (H2III)	1 (1.4 kb)	...	...	...	
Tcr:					
Man:					
C (TCRHUMC)	7 (3.4 kb)				
V (TCRHUMV)	16 (5.9 kb)				
Mouse:					
C (TCRMUSC)	10 (5.0 kb)	...	...	...	
V (TCRMUSV)	25 (9.2 kb)	...	...	...	

Note.—NT = nontranslated; g = germ line; r = recombinant; CONSTHUM = complete C regions of human Ig; INTCONSTHUM = complete introns of C regions of human Ig; VARHUM = complete V regions of human Ig; INTVARHUM = complete introns of V regions of human Ig; CONSTMUR = complete C regions of murine Ig; INTCONSTMUR = complete introns of C regions of murine Ig; VARGERM = complete germ-line V regions of murine Ig; 5VARGERM = 5′-untranslated regions of germ-line V regions of murine Ig; INTVARGERM = complete introns of germ-line V regions of murine Ig; VARREC = complete recombined V regions of murine Ig; INTVARREC = complete introns of recombined V regions of murine Ig; ORIGC = complete C regions of rabbit Ig; ORIGV = complete V regions of rabbit Ig; RATIGC = complete C regions of rat Ig; RATIGV = complete V regions of rat Ig; CHIGC = complete C regions of chicken Ig; CHIGV = complete V regions of chicken Ig; CAIGV = complete V regions of caiman Ig; HLAI = complete heavy chain of HLA I; 5HLAI = 5′-untranslated regions of HLA I heavy chains; INTHLAI = introns of HLA I heavy chains; 3HLAI = 3′-untranslated regions of HLA I heavy chains; HLAII = complete alpha and beta chains of HLA II; 5HLAII = 5′-untranslated regions of HLA II alpha and beta chains; INTHLAII = introns of HLA II alpha and beta chains; 3HLAII = 3′-untranslated regions of HLA II alpha and beta chains; HLAIII = human factor B, HLA III; 5HLAIII = 5′-untranslated regions of human factor B, HLA III; INTHLAIII = introns of human factor B, HLA III; 3HLAIII = 3′-untranslated regions of human factor B, HLA III; H2I = complete heavy chains of H2 I; 5H2I = 5′-untranslated regions of H2 I heavy chains; INTH2I = introns of H2 I, heavy chains; 3H2I = 3′-untranslated regions of H2 I heavy chains; H2II = complete alpha and beta chains of H2 II; 5H2II = 5′-untranslated regions of murine H2 II, alpha and beta chains; INTH2II = introns of murine H2II, alpha and beta chains; 3H2II = 3′-untranslated regions of murine H2 II, alpha and beta chains; H2III = complete B factor of murine H2 III; TCRMUSC = complete C regions of murine T-cell receptors; TCRMUSV = complete V regions of murine T-cell receptors; TCRHUMC = complete C regions of human T-cell receptors; TCRHUMV = complete V regions of human T-cell receptors.

in this sequence and the expected value of NR for that base is $E(NR) = M(N - M + 1)/N$. Likewise, the variance of NR is $V(NR) = [M(N - M + 1)(N - M)(M - 1)]/N^2(N - 1)$. In these equations N is the the length (in bases) of the sequence and M is the number of times a particular base (here A) is found in the sequence.

The value of the test ("NR test" in tables 3, 4) for a given base in each sequence is $NR - E(NR)/\sqrt{V(NR)}$, by comparison with a normal distribution (David and Barton 1962). The test is called NR* when several sequences are tested;

$$NR^* = \frac{\sum_{i=1}^{w} NR_i}{\sqrt{w}},$$

where w is the number of sequences in the workfile.

2. When all four bases are considered together, we use the test NRT, based on the total number of occurrences of homogeneous runs of all lengths, including single bases; for example, if we use the above sequence, the total number of runs (NRT) of any length is nine, A A U G A A A U G C U U A (David and Barton 1962; Gautier et al. 1985). Then the combined statistic ("NRT test" in tables 2, 4) for all four bases is $NRT - E(NRT)/\sqrt{V(NRT)}$.

The expressions for mean and variance of NRT are too complicated to be given here but are given for "homogeneous runs of mononucleotides" in Gautier et al. (1985). Positive values of the NRT test indicate "antirunny" sequences, i.e., those containing a smaller than expected fraction of total bases in runs of length ≥ 2 because too many dispersed bases are present. Negative values identify "runny" sequences, those in which the bases tend to be grouped in homogeneous tracts. The combined statistic for all four bases is

$$NRT^* = \frac{\sum_{i=1}^{w} NRT_i}{\sqrt{w}},$$

where w is the number of sequences in workfile (Gautier et al. 1985). As with NR, the designation NRT applies to a single sequence and NRT* applies to several sequences.

Results

Table 2 presents results obtained with the NRT* test. A value exceeding the statistically significant value (1.96, here rounded off to 2) for the NRT* test indicates runny sequences when the value is negative and avoidance of runs when the value is positive. Most coding Ig sequences (exons) are not significantly runny; they are called nonrunny; that is, the observed amount of bases in runs is close to that expected—and, consequently, nonsignificant values for them appear in table 2. Coding sequences in general have already been found to be less runny than other sequences (Grantham et al. 1985). We observe here the following three groups of runny sequences: (1) introns of IgC and MHC sequences and (2) 5′- and (3) 3′-untranslated ends of MHC genes. Examination of individual sequences in each group (not shown) often reveals significantly elevated amounts of runs. The first two introns of MHC class II genes alpha and beta (both human and murine) are more runny than the succeeding introns (in-

Table 2
Statistical Test on Runs in Sequences Grouped According to Function

Workfile	Sequence Type: Exon	Intron	Switch	5′ End	3′ End
VARGERM	5	NS		2	
VARREC	3	NS			
VARHUM	2	NS			
CONSTMUR	NS	−2			
CONSTHUM	NS	−3			
ORIGC	5				
ORIGV	3				
RATIGC	NS				
RATIGV	NS				
CHIGC	NS				
CHIGV	NS				
CAIGV	NS				
TCRMUSC	−3				
TCRMUSV	NS				
TCRHUMC	NS				
TCRHUMV	NS				
H2I	2	−5		NS	NS
H2II	NS	−8		−4	−5
H2III	NS				
HLAI	2	−9		−3	−2
HLAII	NS	−7		−4	−4
HLAIII	−6	−3			
SWITMUS			8		

NOTE.—See table 1 for sample description by sequence family and species (see Note to table 1 for description of each workfile). For each workfile, the normalized test value appears. Negative values indicate base aggregation, and positive values indicate avoidance of homonucleotides of length ⩾2. This test measures aggregation tendency of all four bases combined. NS = not significant at 5%.

cidentally, the introns in MHC class II genes are not very similar between genes [results not shown]). This difference in sequence-element organization along the sequence recalls observations on another sequence element, CpG (see below), which is unevenly distributed along human and mouse sequences and may be involved in regulation of expression (Max 1984; Tykocinski and Max 1984; Cooper and Gerber-Huber 1985; Grantham 1985; Wolf and Migeon 1985; Grantham et al. 1986).

The four bases are unequally aggregated, as seen in table 3. In immune-system exons, G and C are the most runny bases. In introns, excepting those of the leader-V regions, G and C are always runny while A and T runs are either not statistically significant or are avoided (one exception is seen in table 3). In switch regions runs are strongly avoided in bases A and T.

Thus we recognize the following three kinds of sequences: runny, nonrunny, and antirunny (Grantham et al. 1985, 1986). In table 2, of the 22 workfiles of exons, seven are antirunny, two are runny, and 13 are nonrunny. The most consistently antirunny group of sequences (those showing statistically significant avoidance of runs) seen either in this or in previous work (Grantham et al. 1985) is the workfile SWITMUS, which represents switch regions of mouse JH-CH introns. Why are bases in switch regions dispersed?

Table 3
Statistical Test on Runs of Each Base

Workfile	Run			
	A	C	G	T
VARHUM	5	NS	NS	4
VARGERM	6	3	−4	7
VARREC	4	5	−8	9
CONSTHUM	NS	−3	NS	6
CONSTMUR	NS	−4	NS	6
RATIGC	NS	NS	NS	3
RATIGV	NS	NS	−2	NS
ORIGC	4	NS	NS	5
ORIGV	3	NS	NS	5
CHIGC	NS	−2	NS	NS
CHIGV	NS	NS	NS	NS
CAIGV	NS	NS	−2	NS
TCRMUSC	−3	−2	−3	2
TCRMUSV	NS	NS	−3	3
TCRHUMC	NS	−3	NS	NS
TCRHUMV	NS	NS	−4	3
H2I	4	−3	NS	5
H2II	3	−4	NS	NS
H2III	NS	−2	NS	2
HLAI	6	−2	NS	4
HLAII	4	−3	NS	NS
HLAIII	NS	−3	−2	NS
INTVARHUM	NS	NS	−3	NS
INTVARGERM	NS	NS	−3	3
INTVARREC	NS	NS	NS	2
INTCONSTHUM	4	−5	−6	4
INTCONSTMUR	3	−4	−5	NS
INTH2I	NS	−6	−6	NS
INTH2II	NS	−9	−7	−4
INTHLAI	NS	−9	−8	−2
INTHLAII	NS	−6	−9	−3
INTHLAIII	NS	−4	−3	NS
SWITMUS	8	NS	2	9
5VARGERM	NS	NS	NS	3
5H2I	NS	NS	NS	NS
5HLAI	NS	−2	−3	NS
5HLAII	−4	−2	NS	−3
5H2II	−3	NS	NS	−3
3H2I	NS	−3	NS	NS
3HLAI	−2	−5	NS	NS
3HLAII	−3	−3	NS	NS
3H2II	−6	−4	−3	NS

NOTE.—Data are normalized values for NR* test. Absolute values ≥2 are significant at 5%; negative values indicate runny sequences. See table 2 for NRT* values with all four bases combined on each workfile. NS = no significant difference between observed and expected value.

For each sequence in SWITMUS, table 4 gives the results of tests of statistical significance for runs of each base (NR test) and for all four bases combined (NRT test). We see that nine of the 17 sequences are antirunny in A and that seven are antirunny in T, which is intriguing, since, outside the immune system, A and T have

Table 4
NR and NRT Test on Each Switch Sequence

None	Length (bases)	Switch Region	NR Test A	NR Test C	NR Test G	NR Test T	NRT Test
MUSIGCD08	185	Ig M	3	2	6	3	7
MUSIGCD09	1461	Ig M	4	5	10	8	12
MUSIGCD18	838	Ig G3	2	NS	−2	NS	NS
MUSIGCD19	595	Ig G1	2	NS	NS	NS	NS
MUSIGCD23	262	Ig G2b	NS	NS	NS	NS	NS
MUSIGCD24	1623	Ig G2b	7	−3	NS	5	3
MUSIGCD26	138	Ig G2b	NS	NS	NS	NS	NS
MUSIGCD27	159	Ig G2b	NS	NS	NS	NS	NS
MUSIGCD30	407	Ig G2a	NS	NS	NS	NS	NS
MUSIGCD31	389	Ig G2a	NS	NS	NS	NS	NS
MUSIGCD32	541	Ig G2a	2	NS	NS	NS	NS
MUSIGCD33	165	Ig G2a	NS	NS	−3	3	NS
MUSIGCD34	256	Ig G2a	3	NS	−2	3	NS
MUSIGCD35	418	Ig G2a	2	NS	NS	NS	NS
MUSIGCD36	200	Ig G2a	NS	NS	NS	NS	NS
MUSIGCD39	652	Ig E	NS	4	NS	3	3
MUSIGCD41	1443	IgA	4	4	6	5	8
SWITMUS	9732	Total	8	NS	2	9	8

NOTE.—NS = not significant.

been found to be the most runny bases (Grantham et al. 1985). Many more switch sequences are known for IgG than for other classes, and, curiously, in total run content for all four bases combined, these IgG switch sequences appear more normal than those of IgM, IgE, or IgA. The last three classes avoid runs strongly, with IgA and M switch regions avoiding them in all four bases. These data must be evaluated with caution, however, because sequence length (table 4) varies considerably.

Avoidance of runs implies dispersion of the bases, which could result from repetition of heterogenous oligonucleotides. We find evidence of such repetition in IgM, IgE, and IgA switch sequences (see table 5). Repetitions of longer units (decanucleotides) have also been reported in these sequences, but they derive from the pentanucleotide repetitions (Nikaido et al. 1981, 1982). The sequence MUSIGCD24 (IgG2b switch; table 4) appears antirunny in bases A and T. Although it does not exhibit pentanucleotide repetitions, it does show great internal similarity, since repetitions of blocks of ~80 bases have been observed therein (Kataoka et al. 1981). Thus, IgG may differ from classes M, E, and A in run strategy, since, unlike those of the other three classes most IgG switch sequences are not significantly antirunny in all four bases combined (table 4). However, whether IgG differs from the other classes in run strategy must for now be left unresolved, awaiting further sequencing. Some other switch regions that were analyzed (see table 5) reveal repeated units.

For at least three classes of IgH switch sequences, base dispersion appears to be characteristic (Davis et al. 1980). The heteropolymer repeat units present in these sequences may play a role in the recombination-switching occurring during B-lymphocyte maturation. A test case is the C mu–C delta zone, which apparently contains no switch region 5′ to C delta—and, indeed, the three portions of C mu–C delta reveal no heterogeneous repeat structures and are not antirunny (data not shown).

Table 5
Switch Sequences Containing Tandemly Repeated Heterogeneous Pentanucleotides

Switch Sequence (Class)	Pentanucleotide Repeated (No. of Times), Position
MUSIGCD08 (M)	TGAGC (4), 35 TGAGC (5), 60 TGAGC (5), 90 TGAGC (6), 120
MUSIGCD09 (M)	TGAGC (5), 7 TGAGC (4), 302 TGAGC (4), 349 TGAGC (5), 373 TGAGC (7), 402 TGAGC (5), 472 TGAGC (5), 547 TGAGC (4), 577 TGAGC (5), 632 TGAGC (4), 662 TGAGC (4), 722
MUSIGCD39 (E)	GCTGA (3), 64 GTGCT (3), 103 TGGAC (3), 251
MUSIGCD41 (A)	GCTGA (13), 237
MUSIGHAB4 (mu/alpha)	GGCTG (3), 135 GCTGA (4), 168 GGCTG (3), 388
MUSIGHAI2 (mu/gamma 1)	GGCTG (3), 1505 GGCTG (3), 1591 GGTGA (11), 1622 GAGAG (13), 1695 GGCTG (3), 1805 GGCTG (3), 1933
MUSIGHAN1 (mu/gamma 3)	TGAGC (7), 7 TGAGC (4), 72 TGAGC (5), 97 TGAGC (5), 127 TGAGC (6), 157 TGAGC (3), 192
HUMIGCB2 (mu/alpha)	GCTGA (3), 3 GCTGA (4), 38 TGAGC (3), 80

NOTE.—All repeats are derived from the same type of pentanucleotide repeat, but each time the exact repeated pentanucleotide is given, with the ends of the sequence segment containing the repeats and minor variations in the repeating unit being taken into account.

Possible interaction with translocation is also of interest. Translocations take place between chromosomal target zones and are frequently linked to B-lymphocyte tumors (plasmacytomas, Burkitt lymphoma, etc.). Translocation results from breakage in the c-*myc* locus on human chromosome 8 or mouse chromosome 15, accompanied by breakage in Ig gene sequences (Cory et al. 1983; Rabbitts et al. 1984; Caubet et al. 1985; Corcoran et al. 1985). Table 6 summarizes chromosomal involvements in B-cell lymphoma translocations. These data raise the following two questions: 1. Since,

Table 6
Chromosomal Involvements in B-Cell Lymphoma Translocations

Species	Translocation: c-myc	Ig
Mouse	t(15;12)	H
	t(6;15)	κ
Man	t(8;14)	H
	t(2;8)	κ
	t(8;22)	λ

in the most frequent translocations, the transposed *myc* gene is joined to a switch IgH site but a few translocations (<10%) involve IgL loci, do IgL regions 5′ to C lambda or kappa genes show similarities to switch zones? 2. Does the c-*myc* zone directly implicated in the IgH and IgL translocation contain a repeating heteropolymer like that of the switch regions of IgM, IgE, and IgA?

To answer the first question we have analyzed the J kappa–C kappa zones of a human and a mouse sequence (GenBank release 46). These sequences, the only two published so far, are markedly runny (data not shown); consequently they are opposite in this regard to the antirunny switch sequences.

To respond to the second question we divided human and mouse c-*myc* loci into the following subsequences: 5′- and 3′-untranslated ends, exon 1 (untranslated), exons 2 and 3, and introns 1 and 2. These subsequences have been analyzed using tests NR and NRT. We have also analyzed many oncogen-immunogen sequences. The target zones for translocation are, at least in the mouse, exon 1 and intron 1 or their flanking regions (Stanton et al. 1983; Graham et al. 1985; Klein and Klein 1985). No antirunny zones have been identified either in the c-*myc* region involved in translocation or in the c-*myc*-Ig sequences (data not shown). Therefore we cannot conclude that antirunniness generally characterizes regions implied in recombination events.

Zones in which Z-DNA–to–B-DNA conformation changes (Cantor and Efstratiadis 1984; Ho et al. 1985; Johnston and Rich 1985; Kohwi-Shigematsu and Kohwi 1985) could occur could be hot spots for rearrangements (Hamada et al. 1982; Nordheim and Rich 1983). Analysis of pyrimidine-purine repeats, $(Y\text{-}R)_n$, in target translocation regions of c-*myc* reveals a tendency toward $(Y\text{-}R)_n$ structures (Chorazy 1985), especially in the 5′-untranslated end, in exon 1 and intron 1 of murine c-*myc*. In human c-*myc* the tendency is high in the 5′ NT end, intron 1 and intron 2 (data not shown). However, we have no good information on the extent of the target regions for translocation in human c-*myc*. The 5′ C kappa Ig regions show (Y-R) repeats, whereas switch regions do not present this tendency. Although the repeating pentanucleotides are heteropolymers, they do not give a high concentration of $(Y\text{-}R)_n$ structures.

There is significant evidence for the presence of alternating pyrimidine-purine structures in c-*myc* and 5′ C kappa regions. These structures may be either homologous (divergently related) or analogous (convergently related). However, the $(Y\text{-}R)_n$ structure cannot explain the translocation in general because that structure is not elevated in IgH switch sequences, where 90% of all translocations involving c-*myc* occur. On the other hand, as do other c-*myc* regions, target translocation regions of murine c-*myc*

Table 7
CpG Usage in Different Parts of Murine and Human c-myc Regions

Region	Length (bp)	G + C[a] (%)	No. of CpG	Test Results[b]
Mouse:				
5′ NT	424	60.6	34	NS
Exon 1	565	60.5	37	−3
Intron 1	1,549	57.1	97	−3
Exon 2	754	61.3	55	−3
Exon 3	560	54.3	23	−4
3′ NT	372	33.9	6	NS
Man:				
5′ NT (site 1)	2,327	51.3	102	−6
5′ NT (site 2)	2,501	52.3	118	−5
Exon 1 (site 1)	554	64.4	47	−2
Exon 1 (site 2)	380	64.2	31	NS
Intron 1	1,638	59.5	114	−3
Exon 2	754	64.1	62	−2
Intron 2	1,376	45.9	18	−8
Exon 3	560	52.1	21	−4
3′ NT	866	36.2	4	−6

NOTE.—The second intron of the mouse is not given because its sequence is partial (only 11 bp sequenced). NT = nontranslated; NS = not significant.

[a] Overall G + C content of each sequence.

[b] Sources: Gautier et al. 1985 and Boudraa and Perrin 1987.

genes avoid CG doublets (one example of Y-R structure), contrary to the data of Dunnick et al. (1985). We have tested the observed number of CG doublets against the expected number (Gautier et al. 1985; Boudraa and Perrin 1987). The data in table 7 indicate avoidance (negative test values) of CG doublets in intron 1 and all three exons. Hence, CG doublet avoidance does not appear to be a significant structure for translocations since CpG avoidance occurs throughout the gene, except for the 5′- and 3′-untranslated ends.

Finally, all transposons in the bank were analyzed. In addition, two *Trypanosoma* sequences (Campbell et al. 1984; Kimmel et al. 1985) of special interest have been tested for runs, since antigenic modulation may involve a transposition mechanism similar to Ig switching. We observe that several transposable sequences are antirunny, as are the two *Trypanosoma* sequences.

Discussion

Although the sample is still small, being only ~10 kb, Ig switch sequences have a run content markedly lower than expected. This work is confined to mouse switch regions, since, unfortunately, few human switch sequences are known. The antirunny quality appears to arise mainly from the presence of repeating heteropentanucleotides in the switch regions. Certain exons (particularly V segments) of the immune system also show an antirunny tendency, although the tendency is weaker than in switch zones and V segments exhibit no repetition of heteropolymer units. Dispersion of bases thus appears important in the switching process and may also play a role in translocation.

It is unclear whether the repeating heteropentanucleotides (TGAGC or GCTGA, which are nearly complements) in switch regions (1) induce Z-DNA conformation (although they are not [Y-R] alternations), as seems possible when one considers the work of several other investigators (Cantor and Efstradiatis 1984; Ho et al. 1985; Johnston and Rich 1985; Kohwi-Shigematsu and Kohwi 1985) or (2) have been selected because they increase the complexity of the sequences, thus assuring more highly specific and stable interactions with other sequences. This is a question to be answered by future work. New analyses of target zones for c-*myc*-Ig or c-*myc*-Tcr sequence translocations (Adams and Cory 1985; Bakhshi et al. 1985) will aid in understanding the importance of local sequence structure in these striking genetic processes.

Acknowledgments

We thank C. Gautier, M. Gouy, and T. Greenland for help. We are also grateful for abundant aid from W. Fitch during review of the manuscript, which led us to clarify our methodology. P.P. thanks the Fondation pour la Recherche Médicale for financial support.

LITERATURE CITED

ADAMS, J. M., and S. CORY. 1985. Myc oncogene activation in B and T lymphoid tumours. Proc. R. Soc. Lond. [Biol.] **226**:59–72.

BAKHSHI, A., J. P. JENSEN, P. GOLDMAN, J. J. WRIGHT, O. W. MCBRIDE, A. L. EPSTEIN, and S. J. KORSMEYER. 1985. Cloning the chromosomal breakpoint of t(14;18) human lymphomas: clustering around JH on chromosome 14 and near a transcriptional unit on 18. Cell **41**:899–906.

BOUDRAA, M., and P. PERRIN. 1987. CpG and TpA frequencies in the plant system. Nucleic Acids Res. **15**:5729–5737.

CAMPBELL, D. A., M. P. VAN BREE, and J. C. BOOTHROYD. 1984. The 5′-limit of transposition and upstream barren region of a trypanosome VSG gene: tandem 76 base-pair repeats flanking (TAA)90. Nucleic Acids Res. **12**:2759–2774.

CANTOR, C. R., and A. EFSTRATIADIS. 1984. Possible structures of homopurine-homopyrimidine S1-hypersensitive sites. Nucleic Acids Res. **12**:8059–8072.

CAUBET, J. F., D. MATHIEU-MAHUL, A. BERNHEIM, C. J. LARSEN, and R. BERGER. 1985. Remaniement du locus du proto-oncogène c-myc dans une lignée cellulaire d'origine lymphoblastique T. C. R. Acad. Sci. **300**:171–176.

CHAPLIN, D. D., D. E. WOODS, A. S. WHITEHEAD, G. GOLDBERGER, H. R. COLTEN, and J. G. SEIDMAN. 1983. Molecular map of the murine S region. Proc. Natl. Acad. Sci. USA **80**: 6947–6951.

CHIEN, Y. H., N. R. J. GASCOIGNE, J. KAVALER, N. E. LEE, and M. M. DAVIS. 1984. Somatic recombination in a murine T-cell receptor gene. Nature **312**:322–326.

CHORAZY, M. 1985. Sequence rearrangements and genome instability. J. Cancer Res. Clin. Oncol. **109**:159–172.

COOPER, D. N., and S. GERBER-HUBER. 1985. DNA methylation and CpG suppression. Cell Differ. **17**:199–205.

CORCORAN, L. M., S. CORY, and J. M. ADAMS. 1985. Transposition of the immunoglobulin heavy chain enhancer to the myc oncogene in a murine plasmacytoma. Cell **40**:71–79.

CORY, S., S. GERONDAKIS, and J. M. ADAMS. 1983. Interchromosomal recombination of the cellular oncogene c-myc with the immunoglobulin heavy chain locus in murine plasmacytomas is a reciprocal exchange. EMBO J. **2**:697–703.

CUSHLEY, W., and A. R. WILLIAMSON. 1982. Expression of immunoglobulin genes. Essays Biochem. **18**:1–39.

DAVID, F. N., and D. E. BARTON. 1962. Distribution theory of runs. Pp. 85–101 *in* Combinatorial chance. Griffin, London.

DAVIS, M. M., S. K. KIM, and L. E. HOOD. 1980. DNA sequences mediating class switching in α-immunoglobulins. Science **209**:1360–1365.

DUNNICK, W., J. BAUMGARTNER, L. FRADKIN, and C. SCHULTZ. 1985. DNA sequences involved in the rearrangement and expression of the murine c-myc gene. Curr. Top. Microbiol. Immunol. **113**:154–160.

GAUTIER, C., M. GOUY, and S. LOUAIL. 1985. Non-parametric statistics for nucleic acid sequence study. Biochimie **67**:449–453.

GOUY, M., F. MILLERET, C. MUGNIER, M. JACOBZONE, and C. GAUTIER. 1984. ACNUC: a nucleic acid sequence data base and analysis system. Nucleic Acids Res. **12**:121–127.

GRAHAM, M., J. M. ADAMS, and S. CORY. 1985. Murine T lymphomas with retroviral inserts in the chromosomal 15 locus for plasmacytoma variant translocations. Nature **314**:740–743.

GRANTHAM, R. 1985. CG doublet difficulties in vertebrate DNA (letter). Nature **313**:437.

GRANTHAM, R., T. GREENLAND, S. LOUAIL, D. MOUCHIROUD, J. L. PRATO, M. GOUY, and C. GAUTIER. 1985. Molecular evolution of viruses as seen by nucleic acid sequence study. Bull. Inst. Pasteur **83**:95–148.

GRANTHAM, R., P. PERRIN, and D. MOUCHIROUD. 1986. Patterns in codon usage of different kinds of species. Oxford Surv. Evol. Biol. **3**:48–81.

HAMADA, H., M. G. PETRINO, and T. KAKANAGA. 1982. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Natl. Acad. Sci. USA **79**:6465–6469.

HO, P. S., C. A. FREDERICK, G. J. QUIGLEY, G. A. VAN DER MAREL, J. H. VAN BOOM, A. H.-J. WANG, and A. RICH. 1985. G. T. wobble base-pairing in Z-DNA at 1.0 A atomic resolution: the crystal structure of d(CGCGTG). EMBO J. **4**:3617–3623.

HOOD, L., M. KRONENBERG, and T. HUNKAPILLER. 1985. T cell antigen receptors and the immunoglobulin supergene family. Cell **40**:225–229.

JOHNSTON, B. H., and A. RICH. 1985. Chemical probes of DNA conformation: detection of Z-DNA at nucleotide resolution. Cell **42**:713–724.

KATAOKA, T., T. MIYATA, and T. HONJO. 1981. Repetitive sequences in class-switch recombination regions of immunoglobulin heavy chain genes. Cell **23**:357–368.

KAVALER, J., M. M. DAVIS, and Y. H. CHIEN. 1984. Localization of a T-cell receptor diversity-region element. Nature **310**:421–423.

KIMMEL, B. E., S. SAMSON, J. WU, R. HIRSCHBERG, and L. R. YARBROUGH. 1985. Tubulin genes of the African trypanosome *Trypanosoma brucei rhodesiense:* nucleotide sequence of a 3.7-kb fragment containing genes for alpha and betatubulins. Gene **35**:237–248.

KLEIN, G., and E. KLEIN. 1985. Myc/Ig juxtaposition by chromosomal translocations: some new insights, puzzles and paradoxes. Immunol. Today **6**:208–215.

KOHWI-SHIGEMATSU, T., and Y. KOHWI. 1985. Poly(dG)-poly(dC) sequences, under torsional stress, induce an altered DNA conformation upon neighboring DNA sequences. Cell **43**: 199–206.

MATHER, E. L., K. J. NELSON, J. HAINOVICH, and R. P. PERRY. 1984. Mode of regulation of immunoglobulin μ- and δ-chain expression varies during B-lymphocyte maturation. Cell **36**:329–338.

MAX, E. E. 1984. New twist to DNA methylation (letter). Nature **310**:100.

MORI, L., A. F. LECOQ, F. ROBBIATI, E. BARBANTI, M. RIGHI, F. SINIGAGLIA, F. CLEMENTI, and P. RICCIARDI-CASTAGNOLI. 1985. Rearrangement and expression of the antigen receptor α, β, and γ genes in suppressor antigen-specific T cell lines. EMBO J. **4**:2025–2030.

MÜLLER-EBERHARD, H. J. 1975. Complement. Annu. Rev. Biochem. **44**:697–724.

NIKAIDO, T., S. NAKAI, and T. HONJO. 1981. Switch region of immunoglobulin C gene is composed of simple tandem repetitive sequences. Nature **293**:845–848.

NIKAIDO, T., Y. YAMAWAKI-KATAOKA, and T. HONJO. 1982. Nucleotide sequences of switch regions of immunoglobulin C_ε and C_γ genes and their comparison. J. Biol. Chem. **257**:7322–7329.

NORDHEIM, A., and A. RICH. 1983. Negatively supercoiled simianvirus 40 DNA contains Z-DNA segments within transcriptional enhancer sequences. Nature **303**:674–679.

RABBITTS, T. H., R. BAER, M. DAVIS, A. FORSTER, P. H. HAMLYN, and S. MALCOLM. 1984. The c-myc gene paradox in Burkitt's lymphoma chromosomal translocation. Curr. Top. Microbiol. Immunol. **113**:166–171.

STANTON, L. W., R. WATT, and K. B. MARCU. 1983. Translocation, breakage and truncated transcripts of c-myc oncogene in murine plasmacytomas. Nature **303**:401–406.

TONEGAWA, S. 1983. Somatic generation of antibody diversity. Nature **302**:575–580.

TYKOCINSKI, M. L., and E. E. MAX. 1984. CG dinucleotide clusters in MHC genes and in 5′ demethylated genes. Nucleic Acids Res. **12**:4385–4396.

WOLF, S. F., and B. R. MIGEON. 1985. Clusters of CpG dinucleotides implicated by nuclease hypersensitivity as control elements of housekeeping genes. Nature **314**:467–469.

WALTER M. FITCH, reviewing editor

Received July 7, 1987; revision received October 8, 1987