Proc. Natl. Acad. Sci. USA
Vol. 74, No. 6, pp. 2526–2530, June 1977
Immunology

On the location of palindromes in immunoglobulin genes

(restriction sites/gene translocation/hypervariable regions/antibody diversity)

CHRISTIAN WUILMART*, JACQUES URBAIN†, AND DAVID GIVOL*

* Department of Chemical Immunology, The Weizmann Institute of Science, Rehovot, Israel; and † Laboratory of Animal Physiology, Universite Libre de Bruxelles, 67, rue des Chevaux, B-1640 Rhode-St-Genese, Belgium

Communicated by Michael Sela, February 7, 1977

ABSTRACT **We present a statistical method for detection of palindromes in mRNA or DNA, starting from the protein sequence. Analysis of immunoglobulin genes by this method demonstrates that palindromic sequences are not randomly distributed. They are located at each side of the hypervariable regions in the variable (V) genes, whereas no such regular design is observed in the constant (C) genes. In addition, palindromic sequences overlap the V-C junction in all immunoglobulin classes and significant palindromes are present near residue 216 of the heavy chain, which is the end of deletions in many heavy chain diseases. The relevance of these palindromes to gene translocation and generation of diversity in antibodies is discussed.**

A palindromic DNA sequence possesses a 2-fold rotational axis of symmetry perpendicular to the helix and, as a consequence, the base sequences of the two strands are identical when read in the same polarity:

5′ G-G-G-A-A-A-T-T-T-C-C-C 3′
3′ C-C-C-T-T-T-A-A-A-G-G-G 5′

Such DNA sequences are universally distributed and have been shown to exist in bacteriophages (1, 2), in *Escherichia coli* plasmids (3), and in eukaryotic DNA (4). They can be rather long (300–1200 nucleotide pairs), and some are able to form hairpin-like structures in single-stranded DNA (4). They can also be found in heterogeneous nuclear RNA (5), which is characterized by its resistance to RNase. It has been suggested (6) that their structure (cruciform-like DNA) might be a recognition site for proteins acting at the DNA level; on this basis, molecular models for genetic recombination have also been proposed (7). These speculations are supported by several findings: the *Lac* operator is double stranded and about 27 base pairs long; and the sequence of its RNA transcription copy has been determined (8) and contains a partial 2-fold rotational symmetry. In phage λ, the gene *cI* is surrounded by three left operators and three right operators. Each operator has an axis of partial 2-fold symmetry (9). Inverted repeats have been found in transposable DNA segments of bacteria (10, 11). The specific sites of many restriction enzymes contain short symmetrical sequences, and the palindromic structure allows the cleavage of both DNA strands of opposite polarity at equivalent points (12).

Restriction enzymes have not yet been found in eukaryotic cells, although the properties of some eukaryotic genes suggest their existence. A general hypothesis in immunology is that the variable (V) and constant (C) parts of either heavy or light chains are encoded by separate germ line genes. A translocation mechanism must exist to create contiguous transcribable V-C genes in lymphocyte DNA. The recent data of Hozumi and Tonegawa (13) provide evidence for the existence of such a translocation process. Fougereau and his collaborators have pointed out that palindromic sequences may occur near the V-C junction and that these sequences could be recognition sites for enzymes involved in the translocation process (14).

In this paper, we present a method for detection of palindromes in mRNA, starting from the protein sequences. The method is applied to the case of immunoglobulins, and we show that the palindromic sequences are not randomly distributed but are located preferentially at the boundary of hypervariable regions and near the V-C junction.

Abbreviations: V and C region, variable and constant region, respectively.

METHODS

The searching procedure is an extension of Fitch's polypeptide comparison method (15): the protein sequence is translated into its degenerate mRNA sequence, which is then analyzed for the presence of palindromes. This analysis is performed by matching, in a sliding process, each subsequence of 11 nucleotides (polarity 5′–3′) with all the possible consecutive subsequences of 11 nucleotides (polarity 3′–5′). For each match, a complementation index and its probability for occurring by chance are computed and the coordinates of the complementary segments are stored.

Complementation Test for Matching Nucleotides. Since the genetic code is degenerate, probabilities of complementation must be taken into account in a search for palindromes starting from the amino-acid sequence. The following symbols are used for degenerate positions: X = U/C/A/G, S = C/A, T = U/A, Y = U/C, Z = C/G, R = A/G, and H = A/C/U. Watson–Crick pairs have a complementation score $s(i,j)$ of 36; all other $s(i,j)$ values vary between 0 and 18 according to the probability of complementation.

Matching Experiments of Nucleotide Sequences. The protein sequence is translated into its mRNA sequence according to the genetic code, and the 5′–3′ sequence (A) is compared to itself read in opposite polarity (B). To do this, both A and B are subdivided into spans of 11 contiguous nucleotides; for a sequence of n nucleotides the number of such spans will be $n - 10$. Each span from A is then matched with all the spans of B and complementation analysis is done on the basis of the complementation probabilities. The complementation index, S, is obtained by summing the $s(i,j)$ values characterizing each of the 11 nucleotide matches and its probability is evaluated.

Probability Distribution. The probability of S cannot be derived from its frequency distribution since it has been shown that, for such matching experiments, the distribution of S strongly diverges from a Gaussian curve, especially when the upper values of S are taken into account (16).

In our experiments the probabilities of the different S values have been determined by using the multinomial distribution: if $f_i(A)$ and $f_j(B)$ are, respectively, the frequencies of nucleotide i in sequence A and of nucleotide j in sequence B, the probability for matching i and j is $f_i(A) \times f_j(B)$. Therefore, the probability for obtaining a complementation score value, s, when two randomly selected nucleotides are matched is:

$$p_s = \sum_{s(i,j)=s} f_i(A) \times f_j(B)$$

If two randomly selected spans of 11 nucleotides are matched, the results of 11 matches are assumed to be statistically independent (see below). We wish to compute the probability $p_S^{(11)}$ (S = 0, . . . , 396) for obtaining by chance a complementation index S. The answer is conveniently expressed with the aid of a mathematical auxiliary polynomial function $G(x)$:

$$G(x) = \sum_{s=0}^{36} p_s x^s \qquad [1]$$

For 11 nucleotides the probability $p_s^{(11)}$ is given by the coefficient of x^s in [2]:

$$[G(x)]^{11} = \left(\sum_{s=0}^{36} p_s x^s\right)^{11} \qquad [2]$$

The probability PRO for obtaining by chance a score greater than or equal to a given result S can be defined by Eq. 3:

$$\mathrm{PRO}(S) = \sum_{s=S}^{396} p_s^{(11)} \qquad [3]$$

This distribution of S has been tested by generating 200 Monte Carlo nucleotide sequences with the same length and composition as the flavodoxin-derived mRNA (17), as well as by generating 200 Monte Carlo amino-acid sequences with the same length and composition as flavodoxin and analyzing each of their derived mRNA. Fig. 1 shows that the Monte Carlo and the theoretical distributions match almost perfectly; the deviation observed for probabilities lower than 10^{-6} is due to the relatively small number of Monte Carlo trials. Moreover, the genetic code structure (the third base of the codon is almost always degenerate, while its first and second bases are almost always defined) does not affect the results since both Monte Carlo trials exhibit identical frequency distributions.

Comparison Matrix. The data are more easily handled with a comparison matrix that can give the topology of the palindromes. Therefore the two nucleotide sequences *A* and *B*, which are compared, are recorded along the sides of the matrix. All the spans of 11 nucleotides are matched. For each match, the complementation index, S, is determined and its probability for occurring by chance is derived from the probability distribution of S (see Fig. 1). The probability value PRO(S), corresponding to each match of 11 nucleotides, is put in the matrix at the intersection of the coordinates of the central nucleotide of the two compared spans. To simplify, the probabilities are replaced by numbers ranging from 1 to 4 in the following way: Blank: PRO > 10^{-2}; 1: $10^{-2} \geq$ PRO > 5×10^{-3}; 2: $5 \times 10^{-3} \geq$ PRO > 10^{-3}; 3: $10^{-3} \geq$ PRO > 5×10^{-4}; and 4: $5 \times 10^{-4} \geq$ PRO. After completion, such a matrix is symmetrical and the palindromic sequences are revealed by the occurrence of diagonals with low probability numbers (e.g., 3 and 4) that are perpendicular to the symmetry axis of the matrix.

RESULTS

The Variable Regions. The following 11 human variable region sequences (one per subgroup) have been screened: Eu, V_κI; Ti, V_κII; Mil, V_κIII; New, V_λI; Vil, V_λII; Kearn, V_λIII; Bo, V_λIV; Sh, V_λV; Eu, V_HI; He, V_HII; and Nie, V_HIII (18). For each of them, a comparison matrix has been completed. For example, Fig. 2 gives the COOH-terminal part of V_κII(Ti) (residues 79–109), including the third hypervariable region (19). It is obvious that on both sides of this hypervariable segment there are significant palindromes. Considering the protein sequence, their probabilities for occurring by chance are 1.9×10^{-4} and 0.7×10^{-4}.

Such a localization for the palindromes does not seem to be fortuitous, since analyses of all V_L and V_H subgroups demonstrate palindromes in similar positions on both sides of the third hypervariable region. Moreover, the analysis shows palindromic sequences also on both sides of the first and second hypervariable regions (Fig. 3). The results obtained by analyzing the different subgroups are presented in Table 1; their probabilities of occurring by chance are in the range 10^{-3} to 10^{-4}. Fig. 4 shows the superposition of mRNA sequence on the amino-acid variability plot of Wu and Kabat (19). This combined plot demonstrates that the palindromes exist on both sides of each of the three hypervariable regions but do not overlap hypervariable regions. Table 1 indicates that these palindromes are a constant feature of all immunoglobulin V_L and V_H regions. The localization of these palindromes and their above-mentioned properties suggest that they are being selected for and exist in all V regions due to some important function. The significance of this observation is strengthened by the fact that in none of the constant domains do palindromes appear in a regular way in such positions.

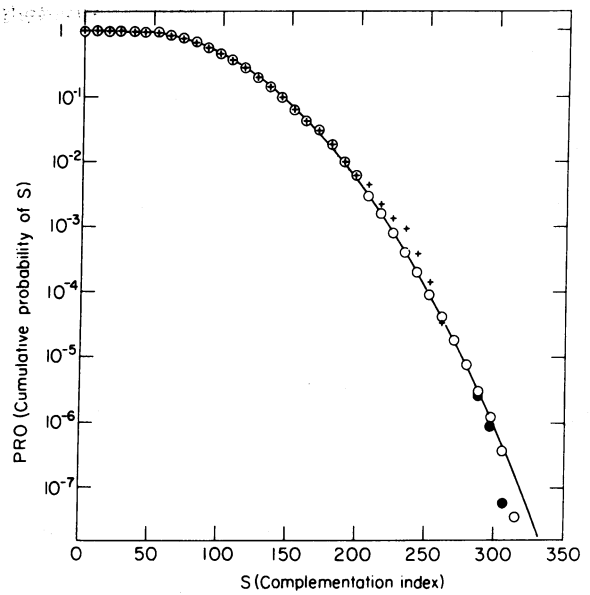

FIG. 1. Cumulative probability distribution of complementation index in the analysis of flavodoxin. (—) Distribution predicted by Eq. 3; (+) analysis of real sequence; (O) observed results for shuffled mRNA sequence; (●) observed results for shuffled protein sequences, when deviating from (O). Both Monte Carlo distributions are identical up to a probability value of 10^{-5}.

V-C Junctions. This area was analyzed by matching the nucleotide sequences corresponding to residue 80–130. The results show a palindromic sequence in the V-C junction which is present in all subgroups analyzed. Part of this palindrome overlaps the end of the V region and part is overlapping up to 30 nucleotides of the C region. This is shown in the following sequences for light chain Ag(κ_I):

103 110
Lys - Leu - Glu - Ile - Lys - Arg - Thr - Val - Ala - Ala - Pro - Ser - Val -
AAR YUX GAR AUH AAR SGX ACX GUX GCX GCX CCX TZX GUX
120 *
Phe - Ile - Phe - Pro - Pro
UUY AUH UUY CCX CCX

and for heavy chain Eu($V_HI\gamma_1$):

110 120
Gly - Gly - Leu - Val - Thr - Val - Ser - Ser - Ala - Ser - Thr - Lys - Gly -
GGX GGX YUX GUX ACX GUX TZX TZX GCX TZX ACX AAR GGX
* *
Pro- Ser- Val- Phe- Pro
CCX TZX GUX UUY CCX

There is a difference between the V-C junctions of the light

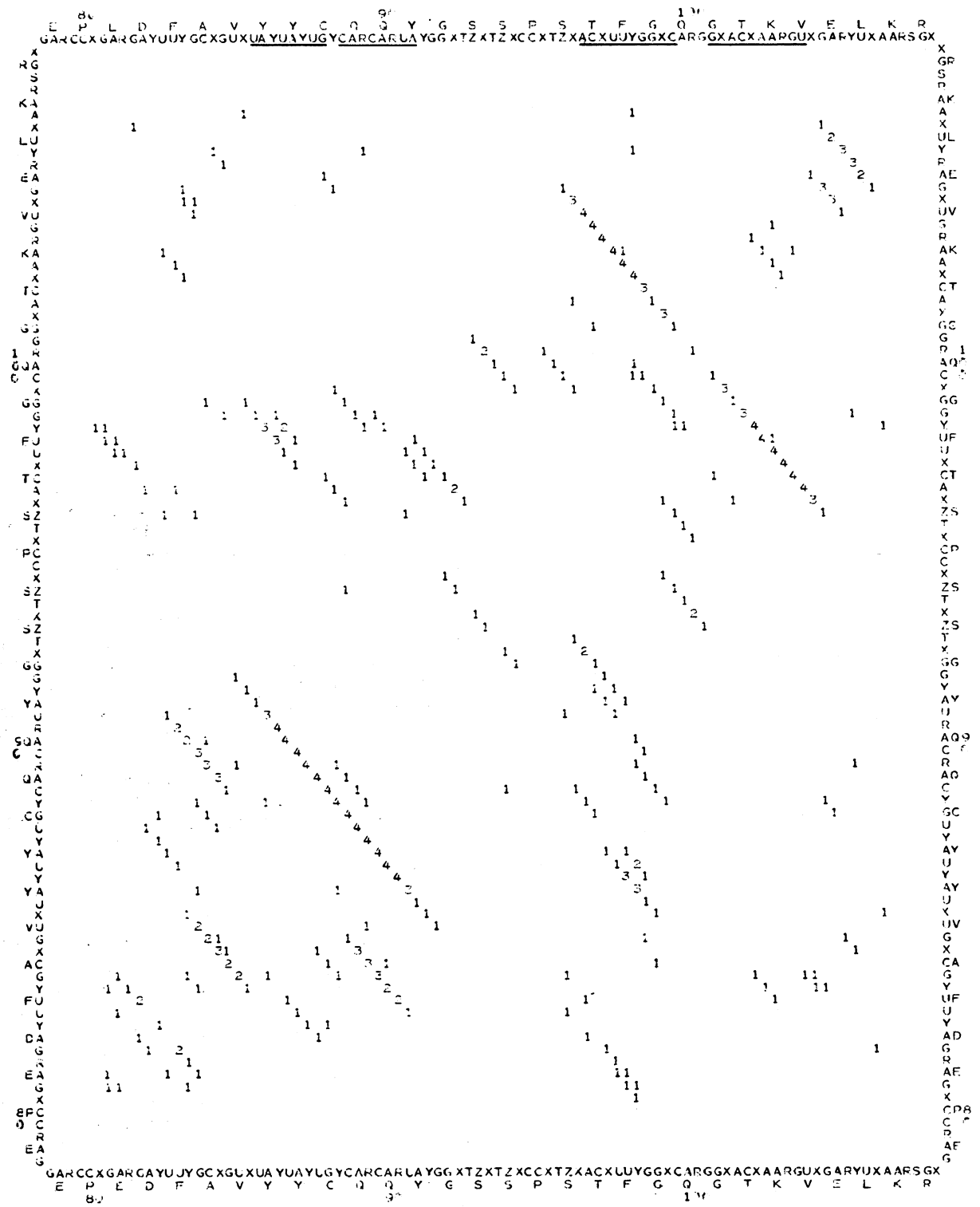

FIG. 2. Comparison matrix obtained by analyzing the COOH-terminal 30 amino acids of V_κII(Ti). The 5′–3′ degenerate mRNA sequence and its inverted form (3′–5′) are reported, respectively, along the abscissa and the ordinate. Palindromic sequences are revealed by the occurrence of diagonals with a high content of numbers 3 and 4 (see *Methods*).

and heavy chains: a single palindromic sequence is found in the light chain while double palindromes are found in the heavy chain. One may speculate that this is related to the fact that V_H regions are shared by all heavy chain classes (γ,μ,α, etc.) whereas V_L regions are confined to one L chain class (λ or κ).

Constant Region. The palindromes in the C region were found to be less significant and, except in the case of the hinge region, they seem to be randomly distributed. They are mainly located around the first cysteine of the intradomain disulfide bridge. Contrary to the V-C junction, no significant palindromes are found in the interdomain junction of Cγ2–Cγ3. On the other hand, the hinge region [in the γ class, it encompasses residues located between positions 221 and 233 (18)] presents

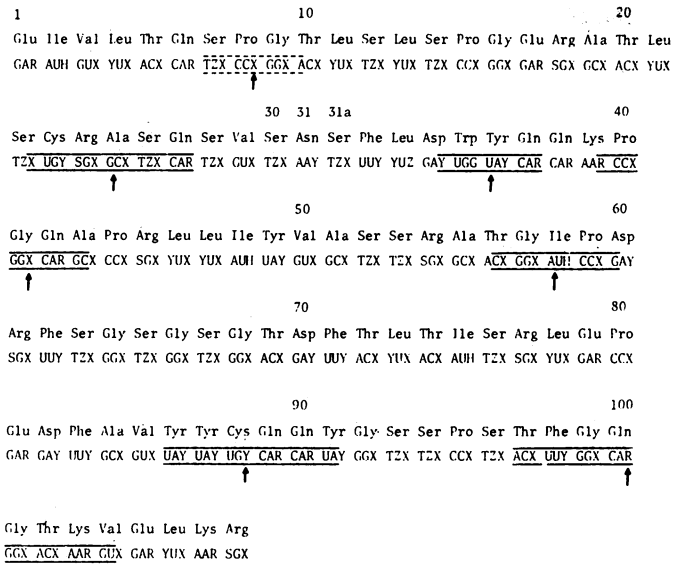

FIG. 3. The amino-acid sequence of V_κII(Ti) is given together with its corresponding degenerate mRNA sequence. Palindromic subsequences are included between lines. Arrows indicate symmetry axis of palindromes.

palindromic sequences running from positions 205–215 and 231–237:

His 204
... GUXAAYCAYAARCCXTZXAAYACXAARGUXGAYAARSGXGUX
Glu 216
GARCCX ... UGYCCXGCXCCXGARYUXYUXGGXGGXCCXTZX ..
Pro 230 Pro 238

Table 1. Distribution of palindromes in human V regions*

V	Hypervariable region I	II	III
V_κI	T22(3)-Q27(3)	L46(3)-Y49(3)	C88(2)-S93(1)
	A34(1)-Q38(2)	S60(2)-G64(3)	G99(3)-E105(3)
V_κII	S22(3)-Q27(3)	K40(3)-A44(2)	Y87(1)-Y92(2)
	N35(3)-Q38(3)	T57(2)-D61(1)	T98(1)-V105(2)
V_κIII	S22(3)-Q27(3)	G45(1)-P48(2)	M93(1)-Q97(3)
	W39(2)-P44(2)	P63(3)-F66(3)	P99(1)-G104(2)
V_λI	P14(1)-G24(2)	L47(2)-D52(2)	A85(1)-A90(2)
	V34(3)-Q38(3)	S57(1)-D61(2)	N97(1)-F101(2)
V_λII	C22(1)-T25(3)	K47(3)-S51(3)	A86(3)-C90(3)
	N33(1)-F38(2)	R56(1)-V60(2)	N97(1)-F101(2)
V_λIII	V18(1)-N25(3)	V45(1)-H48(3)	Y84(2)-W89(1)
	V31(2)-R37(1)	I56(2)-S63(1)	I95(1)-L102(2)
V_λIV	C22(1)-T25(3)	P46(2)-E52(1)	D87(2)-C90(1)
	V35(3)-Q40(3)	P57(1)-R63(2)	N98(1)-L107(2)
V_λV	I19(1)-D24(3)	V45(3)-G48(1)	A83(3)-R89(2)
	W33(1)-Q35(3)	S54(3)-P57(3)	H95(1)-L97(1)
V_HI	V18(2)-S30(1)	V37(3)-W47(2)	D90(2)-Y94(1)
	V37(3)-W47(2)	F55(1)-N59(2)	G101(1)-P105(2)
V_HII	L11(2)-T31(1)	A46(2)-W49(2)	N87(1)-Y97(1)
	V39(3)-R45(3)	D58(1)-P64(2)	P103(3)-D109(2)
V_HIII	C22(1)-T28(3)	A49(3)-S52(3)	B89(2)-C95(1)
	V37(2)-K43(3)	K58(2)-A61(3)	D101(2)-F105(1)

* The table gives the location of palindromes on both sides of hypervariable regions (I, II, and III). Residues are according to the one-letter code. The numbers in parentheses are the position in the codon where palindromes start or end. In addition to these palindromes, another one was found only in the V_κ subgroups between residue Ser 7 and Thr 10.

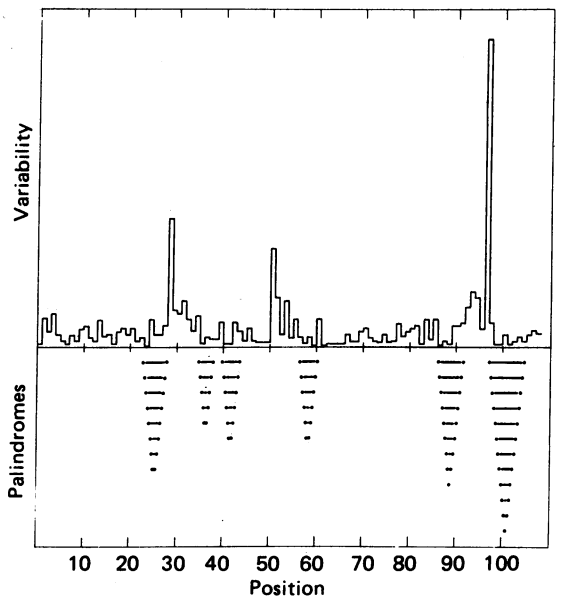

FIG. 4. Location of palindromes with respect to hypervariable regions in V region of light chain. The degenerate mRNA sequence of V_κII(Ti) is superimposed on the variability plot of Wu and Kabat (19). Palindromes are schematically represented in the positions described in Table 1 according to their size. (●—●) Base pairs.

Their probabilities of occurring by chance are, respectively, 2.1×10^{-4} and 3.0×10^{-4}. In this context, it is useful to mention that many heavy chain diseases have big deletions that end at position 216; this is the case with proteins Zuc (20), CRA (21), and Gif (22). Moreover, some heavy chain disease proteins like McG (23) or LEC (24) have a deletion of residues 216–230. The presence of a palindrome before Glu 216 and after Pro 230 supports the suggestion (25) that such proteins might be the result of some unusual repair of broken DNA. Hence, an endonuclease recognizing the above-mentioned palindromes could be a good candidate for the origin of the heavy chain in the disease.

DISCUSSION

The work presented here uses a statistical method to detect reversed repeats in the DNA sequence of immunoglobulin genes, starting from the protein sequence. The analysis was carried out on representative sequences of all V_L and V_H human subgroups, as well as on the constant sequence of $\gamma 1$. The results show significant differences between the predicted DNA structure of V and C genes.

(*i*) In the V genes all three hypervariable regions (19, 26) are bounded by two palindromes on each side, whereas the C genes do not show a comparable distribution of palindromes.

(*ii*) The V-C junction always shows a palindromic sequence that overlaps the end of V and the beginning of C genes, whereas the interdomain area between C_H2 and C_H3 does not show such palindromes. On the other hand, the hinge region shows two palindromic sequences, one between positions 204–216 and one between positions 230–238. Because of the genetic code degeneracy, our computation only detects the possibility for the existence of palindromes. When the sequence of mRNA becomes known, the probability of having palindromes will either increase or decrease, but computation of the S value of individual palindromes shows that even if the degenerate positions do not pair, the PRO value will not be higher than 10^{-2}. It is difficult to dismiss the argument that the presence of these palindromes is a matter of chance. However, their location at nonrandom positions suggests that they have bio-

logical function. Inverse repeats in DNA can be recognition sites for enzymes or preferred sites for recombination; we shall discuss their relevance to the immunoglobulin genes.

The two genes-one polypeptide chain hypothesis already predicted that joining of V and C genes is a necessary step in expression of Ig genes (18). The recent discovery by Hozumi and Tonegawa that DNA fragments that hybridize with V or C mRNA probes are present in different pieces in the embryo but in the same piece in the plasmacytoma DNA (13) provides strong direct support for this hypothesis. The palindrome in the V-C junction is a good candidate for a recognition site by endonucleases that might be involved in V gene translocation and joining to C genes, as was also pointed out by Fougereau *et al.* (14). Similarly, the palindrome around position 216 of the γ chain is compatible with this residue being at the end of many deletions in heavy chains (27).

Therefore, the presence of palindromic structures at the boundary of each hypervariable region suggests that these regions, which are the expression of antibody diversity, may also be related to the action of endonucleases. This implies that somatic or germ line point mutations are not the only mechanisms for explaining antibody diversity and supports other aspects of theories on antibody diversity as follows:

(*i*) It provides support for somatic theories based on the action of specific endonucleases, such as the hypothesis of Brenner and Milstein (28) on diversification as a consequence of DNA breakage and repair. Similarly, the hypothesis of Gally and Edelman (29), which combines gene translocation with intrastrand recombination leading to diversification, is also compatible with these palindromes but restricts the sites of crossing-over of V genes to regions near the hypervariable segments.

(*ii*) The hypothesis of Wu and Kabat (19) about insertion of episomes coding for the hypervariable regions into framework genes [it has been extended by Capra and Kehoe (30) to explain the sharing of idiotypes by different frameworks] is also in line with the location of palindromes on both sides of the hypervariable regions.

(*iii*) The recombinations allowed by palindromic structures (31) are not necessarily restricted to the soma and they may also function in the germ line. The gene expansion hypothesis (32) does not explain the diversity of hypervariable segments after rapid gene expansion when a new set of genes, identical to one member of the old set, appears. The palindromic structures that are also preferred sites for recombination (31) could promote exchange of hypervariable segments between the new and old gene sets. This can conserve variability during V gene expansion (32) and may also explain the sharing of similar hypervariable regions by antibodies of different framework sequences (30).

In either of the above possibilities the predicted occurrence of palindromes in the V genes of immunoglobulins suggests that they are recognition sites for specific enzymes that may be involved in the generation of antibody diversity. On the basis of possible palindromes in the V-C junction and inside hypervariable regions, a similar suggestion concerning the role of restriction-modification enzymes in antibody diversity was recently made by Leder *et al.* (33).

We thank Profs. R. Thomas and L. Segel for helpful discussions. C. W. is a recipient of a long-term EMBO fellowship, on leave from Laboratory of Animal Physiology, Brussels. J. U. is an established investigator of Belgium F.N.R.S.

The costs of publication of this article were defrayed in part by the payment of page charges from funds made available to support the research which is the subject of the article. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

1. Schaller, H., Voss, H. & Gucker, S. (1969) *J. Mol. Biol.* **44**, 445–458.
2. Forsheit, A. B. & Rau, D. S. (1970) *Proc. Natl. Acad. Sci. USA* **67**, 1534–1541.
3. Sharp, P. A., Cohen, S. N. & Davidson, N. (1973) *J. Mol. Biol.* **75**, 235–255.
4. Wilson, D. A. & Thomas, C. A., Jr. (1974) *J. Mol. Biol.* **84**, 115–144.
5. Ryskov, A. P., Saunders, G. F., Farashyan, V. R. & Georgiev, G. P. (1973) *Biochim. Biophys. Acta* **312**, 152–164.
6. Gierer, A. (1966) *Nature* **212**, 1480–1481.
7. Sobell, H. M. (1972) *Proc. Natl. Acad. Sci. USA* **69**, 2483–2487.
8. Gilbert, W. & Maxam, A. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 3581–3584.
9. Ptashne, M., Backman, K., Humayun, M. Z., Jeffrey, A., Maurer, R., Meyer, B. & Sauer, R. T. (1976) *Science* **194**, 156–161.
10. Reanney, D. (1976) *Bacteriol. Rev.* **40**, 552–590.
11. Cohen, S. N. (1976) *Nature* **263**, 731–738.
12. Roberts, R. J. (1976) *Crit. Rev. Biochem.* **14**, 123–164.
13. Hozumi, N. & Tonegawa, S. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 3628–3632.
14. Fougereau, M., Bourgois, A., de Preval, C., Rocca-Serra, J. & Schiff, C. (1976) *Ann. Immunol.* (Inst. Pasteur), **127c**, 607–631.
15. Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
16. McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424.
17. Takata, M., Hanui, M., Yasunobu, K. T., Mayhew, S. & Massey, V. (1971) *Biochem. Biophys. Res. Commun.* **44**, 886–892.
18. Gally, J. A. & Edelman, G. M. (1972) *Annu. Rev. Genet.* **5**, 1–46.
19. Wu, T. T. & Kabat, E. A. (1970) *J. Exp. Med.* **132**, 211–250.
20. Prahl, J. W. (1967) *Nature* **215**, 1386–1387.
21. Franklin, E. C. & Frangione, B. (1971) *Proc. Natl. Acad. Sci. USA* **68**, 187–191.
22. Cooper, S. M., Franklin, E. C. & Frangione, B. (1972) *Science* **176**, 187–189.
23. Fett, J. W., Deutsch, H. F. & Smithies, O. (1973) *Immunochemistry* **10**, 115–118.
24. Rivat, C., Schiff, C., Rivat, L., Ropartz, C. & Fougereau, M. (1976) *Eur. J. Immunol.* **6**, 545–551.
25. Smithies, O., Gibson, J. M., Fanning, E. M., Percy, M. E., Parr, D. M. & Connell, G. E. (1971) *Science* **172**, 574–577.
26. Kabat, E. A. & Wu, T. T. (1971) *Ann. N.Y. Acad. Sci.* **190**, 382–393.
27. Franklin, E. C. & Frangione, B. (1975) in *Contemporary Topics in Molecular Immunology*, eds. Inman, F. P. & Many, W. J. (Plenum Press, New York), Vol. 4, pp. 89–124.
28. Brenner, S. & Milstein, C. (1966) *Nature* **211**, 242–243.
29. Gally, J. A. & Edelman, G. M. (1970) *Nature* **227**, 341–348.
30. Capra, J. D. & Kehoe, J. M. (1975) *Adv. Immunol.* **20**, 1–40.
31. Wagner, R. E. & Radman, M. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 3619–3622.
32. Hood, L., Campbell, J. H. & Elgin, S. C. R. (1975) *Annu. Rev. Genet.* **9**, 305–353.
33. Leder, P., Honjo, T., Seidman, J. & Swan, D. (1976) *Cold Spring Harbor Symp. Quant. Biol.* **41**, 781–788.