Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 1218–1222, February 1990
Immunology

Various regulatory sequences are deprived of their uniqueness by the universal rule of TA/CG deficiency and TG/CT excess

(palindromes/recombination signals/enhancers/plant genes)

SUSUMU OHNO AND TETSUYA YOMO

Beckman Research Institute of the City of Hope, 1450 East Duarte Road, Duarte, CA 91010-0269

Contributed by Susumu Ohno, October 20, 1989

ABSTRACT **The universal rule of TA/CG deficiency–TG/CT excess endures the extremely high mutation rate of a retrovirus (human immunodeficiency virus type 1) as well as methylation of CAG rather than CG in a plant (maize). Among the consistently abundant nucleotide oligomers, there are two complementary pairs of palindromic nucleotide pentamers containing TG and CA. Out of the CAGTG and CACTG pair emerged the heptameric pair for the long-distance recombination of immunoglobulin genes, CACAGTG and CACTGTG. Reflecting their origin, these heptamers are found everywhere in all DNA, and a substantial fraction of them are accompanied by nonameric components properly spaced from them. It appears that, were the recombination event not confined to B cells, results of illegitimate recombinations might be disastrous. The other pentameric pair is TGCAT and ATGCA. Out of this pair emerged the complementary pair of transcription enhancer decamers: TNATTTGCAT for immunoglobulin light chains and ATGCAAATNA for immunoglobulin heavy chains. Again reflecting their origin, these decamers are found everywhere in all DNA and some genes—for example, in the 3′ flanking region of immunoglobulin heavy chain constant region—are accompanied by a downstream "TATA box." It seems that even with regard to the productively recombined immunoglobulin genes, misinitiation of enhanced transcription is a real possibility.**

The universal rule of TA/CG deficiency–TG/CT excess was originally proposed as the construction principle of all coding sequences (1), and this rule was subsequently found to apply to the noncoding regions of genes as well (2). In this paper, the universality of this rule was tested on one viral genome which has been changing at the phenomenal rate of 10^{-3} per base pair per year (3) as well as on a large plant gene in which methylation involves C of CAG rather than C of CG (4). Thus, in plant genes, consistent deficiency of CG dimer cannot readily be attributed to constant conversion of methylated CG to TG and CA. After verification of this rule as the catholic principle, interesting features of several palindromic base oligomers are discussed in connection with their roles as regulatory signals.

Genome of Human Immunodeficiency Virus Type 1 (HIV-1)

Reverse transcriptase of retroviruses is the most error-prone of various nucleic acid polymerases. Its error rate has been estimated as of the order of 10^{-3} per base pair per year, which is roughly 3×10^6 times higher than that of the mammalian DNA polymerase (3). Most retroviruses protect themselves from their own error-prone polymerase by frequent integration into the host genome. In the integrated state, their genome would undergo changes at the host rate of 3×10^{-9} per base pair per year. Not so with HIV-1, the causative agent of human acquired immunodeficiency syndrome (AIDS). As it quickly destroys its preferred target cell type, helper T cells, it has been condemned to evolve at the mercy of its own extremely error-prone reverse transcriptase. As the reverse transcriptase itself must necessarily be mutating at the above-noted rate, its error rate too must be shifting constantly. At any rate, the published complete sequence of the HIV-1 genome consists of 9718 nucleotides (5). Treated as the double-stranded DNA, it is an A+T-rich sequence, 58% A+T. As a single-stranded RNA, treated as DNA in Fig. 1, it is extremely A-rich (35%) and C-poor (18%). Nevertheless, as shown in Fig. 1, the TA/CG deficiency–TG/CT excess rule holds true in the HIV-1 viral genome as well. CG deficiency was extremely severe, while TA deficiency was very moderate, yet significant because of the large number of bases involved in this analysis. The deficiency of these two dimers was, as always, compensated by excesses to nearly the same degree of three nucleotide dimers: TG, CT, and CA. Because of the extreme A-richness of this genome, the top four nucleotide trimers were AAA, AGA, AAG, and GAA. Yet, in terms of the observed/expected, CAG, ranking fifth, was the most overrepresented nucleotide trimer, its complementary nucleotide trimer, CTG, also being in excess to the similar degree. The most scarce were four complementary pairs of nucleotide trimers containing CG. Thus, the universal rule did prevail even under the trying circumstances of an extremely high mutation rate.

Sucrose Synthase Gene of Maize

The coding region of this long plant gene is comprised of 11 exons, 802 codons altogether. The sequenced noncoding region of this gene is nearly twice as long as the coding region, due to the inclusion of 2832 nucleotides in the 5′ flanking region (6). Observing Fig. 2, it should be noted that this gene is hardly distinguishable from mammalian genes previously analyzed (2). Base compositions are again very different between coding and noncoding regions. The former was G+C-rich (53%), whereas the latter became A+T-rich (56%). TA deficiency was more pronounced in the coding region, whereas CG deficiency was more severe in the noncoding region. Yet, three dimers, TG, CT, and CA, were excessive to nearly the same degrees in both regions. The most numerous of the 64 nucleotide trimers was CTG in the coding region, and its complementary trimer, CAG, was also well represented. Because of the noncoding region's extreme T-richness (34%), CTG in this region ranked 3rd, being superseded by TTT and CTT. Yet, in terms of the observed/expected, CTG was the most overrepresented nucleotide trimer in the noncoding region and its complementary trimer, CAG, was also overrepresented. Four complementary pairs

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: HIV-1, human immunodeficiency virus type 1; V_H, heavy chain variable region; C_H, heavy chain constant region; D_H, heavy chain diversity region.

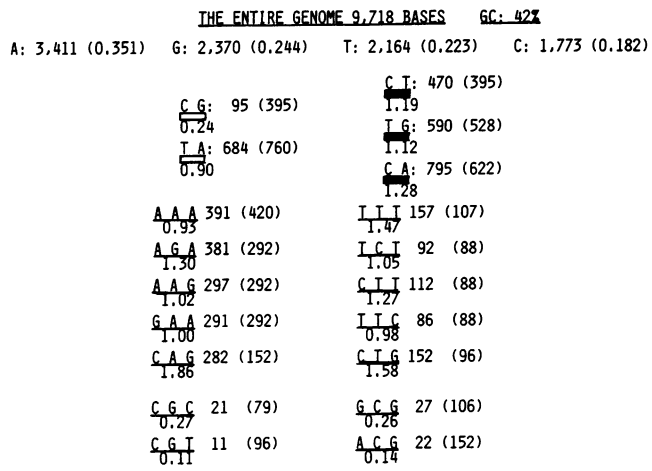

FIG. 1. Dimer and trimer analysis of the HIV-1 retroviral genome, which consists of 9718 nucleotides and is 42% G+C (5). No distinction was made between coding and noncoding regions. At the top, the genome's base composition is shown in numbers and fractions of the four bases. Immediately below are the excessive and deficient dimers. The observed number of each is contrasted with its expected number in parentheses. The observed/expected value is shown immediately below each dimer. Excessive dimers are underlined by solid bars and deficient dimers by open bars. As to nucleotide trimers, the five most abundant trimers are shown at the top of the left column, followed by the two least numerous trimers. In the right column, incidences (observed and, in parentheses, expected as well as observed/expected) of nucleotide trimers complementary to those of the left column are shown.

of TA-containing nucleotide trimers were at the bottom in the coding region, whereas this honor went to four pairs of CG-containing nucleotide trimers in the noncoding region. It appears that the universal rule of TA/CG deficiency-TG/CT excess is ancient in origin, being over and beyond the methylation mechanism involving C of CG. It should be remembered that methylation in plants involves C of CAG rather than of CG (4).

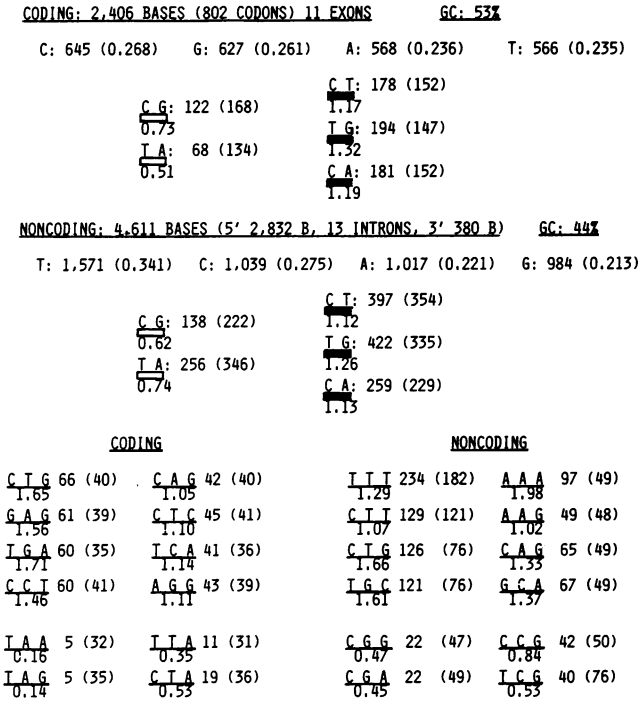

FIG. 2. Dimer and trimer analysis of the sucrose synthase gene of maize (6). The 2406-base-long coding region and the 4611-base-long noncoding region are treated as separate entities. As to nucleotide trimers, only the top four and bottom two of each region and their complementary trimers are dealt with.

Immunoglobulin μ Heavy Chain Constant Region Gene of an Insectivore

The first placental mammals to emerge from the shadow of the dinosaurs 70 million years or so ago were insectivores. Thus, insectivores are equally related to all other placental mammals. For this reason, we show in Fig. 3, as an additional example of mammalian genes, an analysis of the μ C_H gene of the insectivore *Suncus murinus* (7). The coding region and the sequenced noncoding region are nearly equal in length, which was very convenient. Again, there is a 7% difference in G+C content between the coding and noncoding regions. Yet the noncoding region did not become A+T-rich, still maintaining 51% G+C. For this reason, CTG became the most numerous nucleotide trimer of both regions, and the absolute numbers of CTG and its complementary CAG were nearly identical in both. Since TA deficiency was more pronounced in the coding region, the least represented of the nucleotide trimers in this region were four complementary pairs of TA-containing nucleotide trimers, immediately above them in frequency being four complementary pairs of CG-containing trimers. In the noncoding region, in comparison, the order of TA-containing and CG-containing trimers was reversed, due to the more pronounced deficiency of CG.

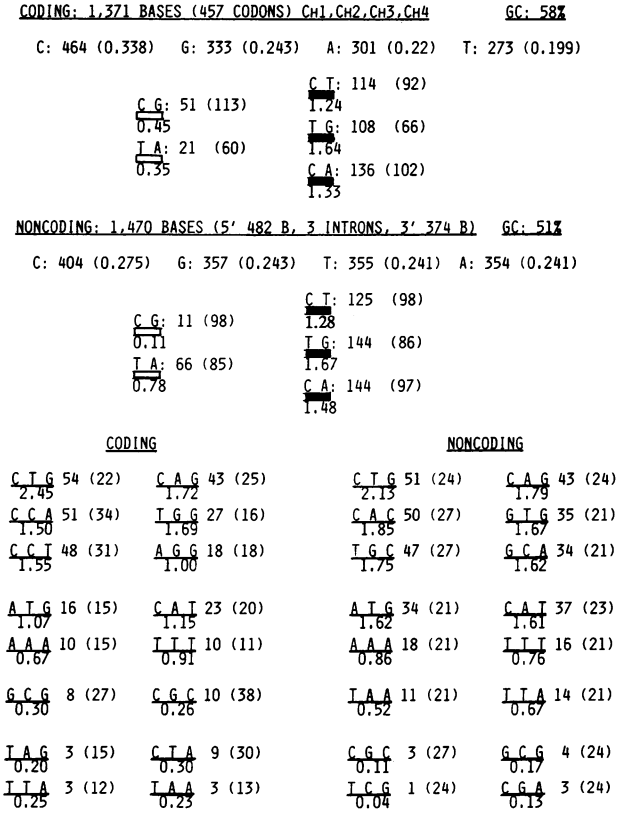

FIG. 3. Dimer and trimer analysis of the immunoglobulin μ C_H gene of the insectivore *Suncus murinus* (7). The 1371-base-long coding region and the 1470-base-long noncoding region are treated as separate entities. As to trimers, in addition to the top three and bottom two trimers and their complementary trimers of each region, ATG and AAA and their complementary trimers CAT and TTT are inserted in the middle as representatives of those that are more numerous in the noncoding region. Also shown, in the third row from the bottom, are the least numerous pair of CG-containing trimers in the coding region as well as the least numerous pair of TA-containing trimers in the noncoding region.

In spite of these minor differences, the noncoding region was merely a scrambled version of the coding region, and because of the imposed symmetry between two strands, each strand was composed of a succession of palindromes.

Pertinent Palindromic Oligonucleotides That Attract DNA-Binding Proteins

Since the operator base sequence of the *lac*-operon system of *Escherichia coli* was determined, the preferential recognition by DNA-binding proteins of palindromic sequences has been well established. This preference by DNA-binding protein is readily understandable, for only by recognizing a palindromic base oligomer can a protein perceive a DNA segment as a single entity, albeit double-stranded. In a perfect palindrome, both strands present the identical 5′ to 3′ base sequence. When confronted with palindromic sequences, regulatory proteins are reasonably discriminating. Yet they are quite inept in discerning the complementarity between a pair of nonpalindromic base oligomers. The above-noted trait is exemplified by the recombinase that mediates the long-distance somatic recombination events involving immunoglobulin genes. Three chromosomes in B lymphocytes are involved in such events. Below we discuss one of the three, one carrying a large number of genes for immunoglobulin heavy chains. These genes are scattered over a vast area, perhaps over 10^6 base pairs long. Antigen-binding variable (V) regions are mainly encoded by hundreds of V_H genes. Tens of very short diversity (D_H) genes and five slightly longer joining (J_H) genes located 5′ to the μ C_H gene (7) of Fig. 3 also contribute terminal segments to variable regions. For each B lymphocyte and its clone to produce a monoclonal IgM antibody, two recombination events have to take place on this chromosome. First, one of the tens of D_H genes fuses with one of the five J_H genes, discarding a long stretch of intervening DNA as an extra chromosomal ring. The second fusion takes place between one of the hundreds of V_H genes and this already fused D_H + J_H–$C_{\mu H}$ complex, and an even longer stretch of DNA is discarded. Signal base oligomers for all these somatic recombination events involving immunoglobulin genes are the same. They are complementary pairs of nucleotide heptamers and nonamers, the exact required spacing between a heptamer and its attendant nonamer being either 23 or 12 bases (8–10). The recombination process between V_H and D_H is schematically illustrated at the top of Fig. 4. It can be seen that the complementarity between heptamers is perfect, whereas that between nonamers is only 5/9. While the 3′ and 5′ canonical types can readily be assigned for heptameric components, one can discern only vague trends with regard to 3′ versus 5′ nonamers. As to heptameric components, both the 3′ canonical type, CACAGTG, and the 5′ type, CACTGTG, are palindromes. Thus the two are simultaneously complementary (7/7) and homologous (6/7) with each other. With palindromic heptamers, the recombinase has been quite discriminating, not allowing the complementarity to fall beyond 5/7. Nevertheless, CACACTG rather than the canonical CACAGTG is found at the 3′ end of many V_H genes (11). Conversely, 5′ ends of more D_H genes are capped by CACAGTG, TACTGTG, and CGCTGTG rather than by the canonical CACTGTG (12). At any rate, the power of discrimination by the recombinase shows a precipitous decline when dealing with nonpalindromic nonameric components. Hence, nonamers became a polyglot lot. Accordingly, there is only a trend of 3′ nonamers containing more A and C bases, while an opposite trend is shown by 5′ nonamers, which have more T and G bases. Consequently, the usual complementarity between 3′ and 5′ nonamers is only 5/9, as shown at the top of Fig. 4, quite often declining to 4/9.

From the above, it became clear that the cornerstone of the immunoglobulin-specific long-distance recombination signal is its heptameric components: the 3′ CACAGTG and 5′ CACTGTG and their single-base-substituted allies. Contained within the above heptameric pair is a pair of palindromic pentamers, CAGTG and CACTG, that were abundant in all genes, not only because of TG/CA excess but also because of the symmetry between complementary strands enforced by the universal rule (1, 2); the immunoglobulin C_H gene (7) of Fig. 3 contains eight CAGTG and seven CACTG. As with the other complementary pair of pertinent palindromic pentamers to be discussed shortly, these CAGTG and CACTG sequences invariably became parts of longer, hexameric to octameric, oligomers. Two kinds of such complementary pairs of octamers are shown in the middle of Fig. 4. CAGTGCAG and its single-base-deviant, CAGTGCTG, are found in the C_H2 exon (left column) while CTGCACTG, complementary to the above, was found in two copies; one in the C_H3 exon and the other further upstream in the 5′ noncoding region (right column).

The other kind of complementary octamers consisted of TCAGTGTG in the left column and CACACTGA in the right column. These two were actually one-base-added versions of 5′ and 3′ canonical heptamers of the recombination signal, for CAGTGTG and CACACTG differed from the canonical CACTGTG and CACAGTG by only a single base each. Moreover, CACACTG was present at the end of many a V_H gene (11). Going back to the first kind of complementary octamers, CAGTGCTG of the C_H2 exon had its attendant nonamer, CCAAGGGCT, exactly 12 bases downstream of it, and the CTGCACTG octamer of the C_H3 exon also had its attendant nonamer, AGGAATGGG, 23 bases upstream of it. Shown in the middle center of Fig. 4 is a mock recombination between the above-noted octamer–12-base–nonamer and the nonamer–23-base–octamer combination. The 7/8 followed by 5/9 complementarity seen in this complex is as respectable as that seen between components of the immunoglobulin-specific recombination signal. The very fact that such a mock, yet elaborate recombination complex can be produced within a span of 520 bases illustrates the enormity of the problems confronting the immunoglobulin-specific recombination event. Because of the universal rule, heptamers containing CAGTG, CACTG and/or CACAG, and CTGTG are very abundant in all DNA. Even the 94-codon-long mouse Pch 108A V_H gene (11), used as an example at the top of Fig. 4, contained an extra heptamer, TACTGTG, in its coding region. As already noted, this heptamer was found at the 5′ end of a number of D_H genes (12). As to the insectivore C_H gene (7), the CTGCACTG octamer of the C_H3 exon (previously discussed and used for a mock recombination) overlapped with the canonical 5′ heptamer, CACTGTG, of the real recombination signal.

It should come as no surprise then if this immunoglobulin C_H gene contained either the 3′ half or the 5′ half of the real recombination signal. As shown at the bottom of Fig. 4, the 3′ noncoding region of this gene indeed contained the nonamer–12-base–heptamer combination that can pass for the 5′ half of the recombination signal normally associated with D_H genes. Exactly 12 bases upstream of the CACTGTA was the TTGTGTGTG nonamer, which showed 5/9 complementarity with the 3′ signal nonamer of the mouse Pch 108A V_H gene (11) already discussed. The chromosomal region in which V_H and D_H genes are scattered is very long, perhaps approaching 10^6 base pairs, as already noted.

In view of the above discussion, one can only conclude that this long region must contain rather large numbers of illegitimate 3′ and 5′ halves of the recombination signal. It should come as no surprise that the expression of immunoglobulin genes is invariably hemizygous in B-cell clones, since the probability of recombination events succeeding in both ho-

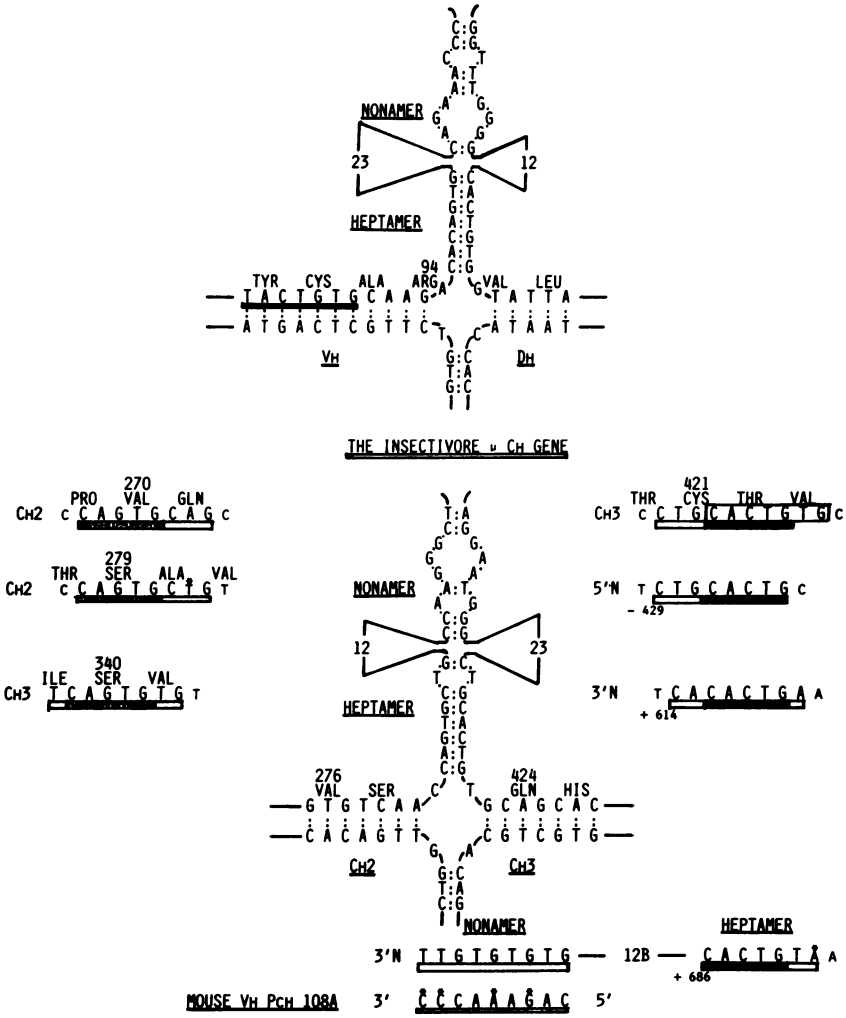

FIG. 4. At the top, the immunoglobulin-specific recombination between one V_H gene, actually mouse Pch 108A (11), and one D_H gene (12) is schematically illustrated. The coding region of mouse Pch 108A V_H contained an extra heptamer, TACTGTG (underlined by a solid bar), which was often found at the 5′ end of D_H genes (12). Shown in the middle are two complementary sets of octamers, found within the 2841-base-long sequence of the insectivore μ C_H gene (7); the ones at the left contained CAGTG (underlined by shaded bars), while the ones at the right included CACTG (underlined by solid bars). The first set contained two members of each. CAGTGCAG and its single-base-substituted copy, CAGTGCTG, both residing in the C_H2 exon, were complementary to CTGCACTG residing in the C_H3 exon. Another CTGCACTG was in the 5′ noncoding region. The second complementarity was between the TCAGTGTG octamer residing in the C_H3 exon and CACACTGA residing in the 3′ noncoding region. The octamer at the top of the right column overlapped with the 5′ canonical heptamer, CACTGTG (boxed), of the immunoglobulin-specific recombination signal. The fact that the signal, which is very much like the immunoglobulin-specific recombination signal, can easily be established between complementary octamers of the first set is illustrated in the center. Shown at the very bottom is the nonamer–heptamer combination in the 3′ noncoding region of this gene, which can be mistaken for the 5′ half of the immunoglobulin-specific recombination signal complex in front of D_H. Shown immediately below the nonameric portion (underlined by the open bar) of this illegitimate 5′ half of the signal is the 3′ signal nonamer of the mouse Pch 108A V_H already introduced at the top (noncomplementary bases are marked by asterisks). Positions of these oligonucleotides in the 5′ and 3′ noncoding regions are indicated by numbers of their first base; e.g., −429 in the 5′ noncoding region means 429 bases upstream of the initiating codon ATG, whereas +614 in the 3′ noncoding region means 614 bases downstream of the last codon.

mologous chromosomes is virtually nil. There must be many a B cell in which all recombination events have failed and aborted. Such is the price to be paid for the regulatory protein's preference for palindromic sequences that are made abundant by the universal rule (1, 2). Since this rule applies to all genes, illegitimate recombination signals must be plentiful in all chromosomes. Our estimate, based on 10 functionally unrelated mammalian genes, is that either 3′ or 5′ recombination signals of an illegitimate nature occur, on the average, once every 2500 bases. It stands to reason that such recombination events take place only in B lymphocytes, and similar events only in T lymphocytes. Involvement of more cell types than these two would have been disastrous.

Another consistently abundant complementary pair of palindromic pentamers was TGCAT and ATGCA, both containing the tetrameric palindrome TGCA. In Fig. 5, various fates of 24 TGCA tetramers contained in the immunoglobulin C_H gene (7) of Fig. 3 are traced. It should be noted that exactly half of them became 6 each of the complementary pair of palindromic pentamers, TGCAT and ATGCA. Out of this pair evolved a decameric pair, TNATTTGCAT and ATGCAAATNA. These two are not only used as 5′ transcription enhancers in concert with the downstream "TATA box" of immunoglobulin light and heavy chain genes (13, 14) but also serve a similar purpose for a number of genes of viruses, prokaryotic as well as eukaryotic (14). Of the six pairs of TGCAT and ATGCA pentamers identified in Fig. 5, one pair was furnished by TGCA tandem repeats shown at the top. Furthermore, the ATGCA portion of this repeat was half of the ATGCAAGTGA decamer that differed only by a single base from the canonical 5′ enhancer decamer of immunoglobulin heavy chain genes. The remaining five pairs of

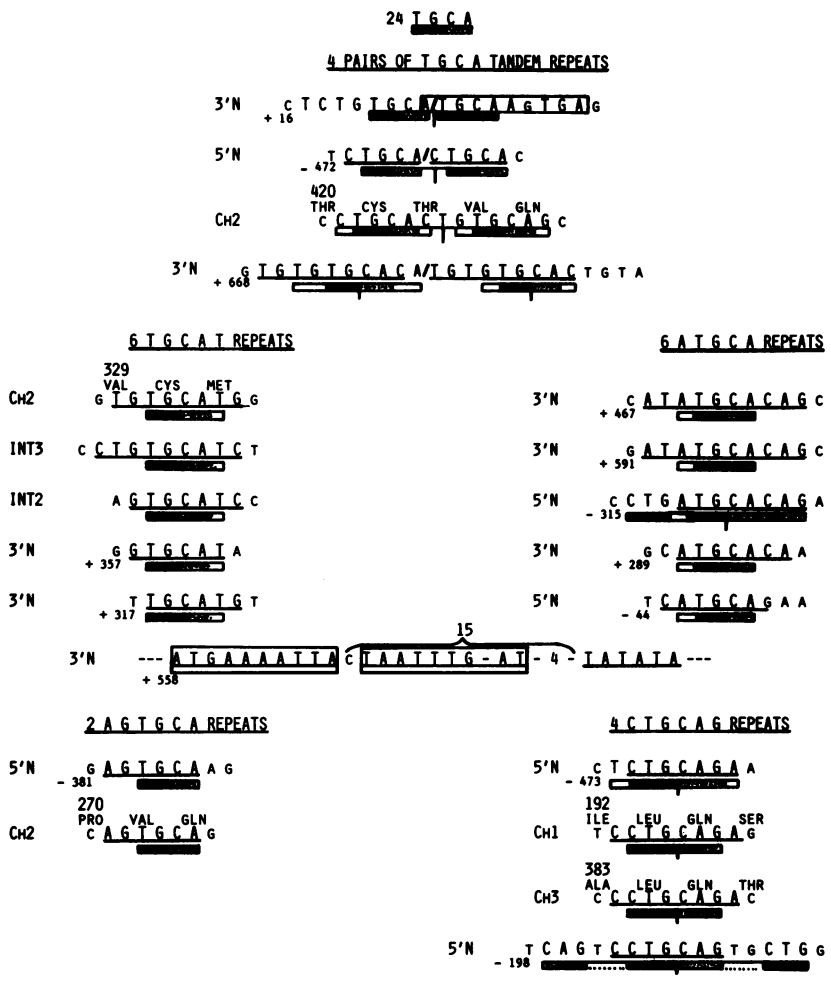

FIG. 5. Various fates of 24 copies of the TGCA palindromic tetramer, also found in the immunoglobulin C_H gene (7) of Fig. 3, are traced. 3'N and 5'N, 3' and 5' noncoding; INT, intron. All of them became parts of pentameric to decameric repeating units (underlined). At the same time, 10 of them became parts of hexameric to hexadecameric palindromes (the center of each palindrome is indicated by the pronounced vertical line). Shown at the top are four pairs of tandem TGCA repeats, two copies being separated by 0, 1, 3, and 6 bases. Two complementary pentamers, TGCAT and ATGCA, overlapped with each other to form the first pair of tandem repeats; the TGCATGCA octamer. The last five bases of this octamer also represented the 5' half of the decamer, ATGCAAGTGA (boxed), which is a single-base-substituted version of the 5' canonical enhancer of immunoglobulin heavy chain genes (13, 14). The remaining five each of TGCAT (left) and ATGCA (right) are shown in the middle. Needless to say, TGCAT is the last half of the 5' canonical enhancer of immunoglobulin light chain genes, whereas ATGCA is the first half of the 5' canonical enhancer of immunoglobulin heavy chain genes. Shown immediately below these pairs of complementary pentamers, TGCAT and ATGCA, is the 30-base-long sequence found in the 3' noncoding region. ATGAAAATTA (underlined by the open bar and boxed) is the proven 5' enhancer of mouse Pch 104 V_H (11), and 14 bases downstream of it is the genuine TATA box, TATATA. Furthermore, separated by 1 base from the above-noted decamer is the TAATTTGAT nonamer (also underlined by the open bar and boxed), which is a single-base-deleted version of the canonical enhancer decamer for immunoglobulin light chain genes.

complementary pentameric palindromes, shown in the middle, generated hexameric to nonameric pairs of complementary oligomers. For example, the CTGTGCATC nonamer in the left column, second row, was complementary to the GATGCACAG nonamer occupying the right column, third row. Similarly, the TGTGCATGC nonamer containing 7/8 of TGCA tandem repeats shown at the top, was complementary to the GCATGCACA nonamer in the fourth row, right column of the middle. The above serve to illustrate that TGCAT and ATGCA invariably become parts of longer complementary oligomers and that each such oligomer tends to recur. For instance, there were two copies of the ATATGCACAG decamer as seen in the first and second rows of the right column in the middle of Fig. 5. It was then inevitable that this immunoglobulin C_H gene (7) should contain at least one enhancer decamer–TATA box complex illegitimately placed. As shown immediately below pairs of TGCAT and ATGCA derivatives in Fig. 5, the 3' noncoding region of this gene indeed contained the ATGAAAATTA decamer, which is a proven enhancer of the mouse Pch 104 V_H gene (11), and 15 bases downstream of this ill-placed decamer was a genuine TATA box, TATATA. In addition, this decamer was but one arm of the 20-base-long palindrome, the other arm being TAATTTG-AT, a single-base-deleted version of the canonical immunoglobulin light chain enhancer decamer.

It is obvious that, because of the TA-deficiency part of the universal rule (1, 2), the TATA box-like sequences are seldom found in the coding region. For example, the coding region, totaling 1371 bases of the immunoglobulin C_H gene (7) of Fig. 3, did not contain a single TATA tetramer. Yet, TA deficiency is always less pronounced in the noncoding region and the TATA box is as ill-defined a hexamer as nonameric portions of the recombination signal; a number of hexameric sequences such as TTAAAT and TTAAAA suffice. For this reason, each enhancer-like decamer in the noncoding region has a good chance of being accompanied by a TATA box, especially since a TATA box can be as far downstream from it as 35 bases or more (13, 14).

It would be of interest to find out to what extent mammalian and other eukaryotic cells may be enduring misinitiated, yet enhanced, transcriptions.

1. Ohno, S. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 9630–9634.
2. Yomo, T. & Ohno, S. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 8452–8456.
3. Gojobori, T. & Yokoyama, S. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4198–4201.
4. Gruenbaum, Y., Naveh-Many, T., Cedar, H. & Razin, A. (1981) *Nature (London)* **292,** 860–862.
5. Rabson, A. B., Daugherty, D. F., Venkatesan, S., Boulukos, K. E., Benn, S. I., Folks, T. M., Feorino, P. L. & Martin, M. (1985) *Science* **229,** 1388–1390.
6. Werr, W., Frommer, W. B., Mass, C. & Starlinger, P. (1985) *EMBO J.* **4,** 1373–1380.
7. Inshiguro, H., Ichihara, Y., Namikawa, T., Nagatsu, T. & Kurosawa, Y. (1989) *FEBS Lett.* **247,** 317–322.
8. Sakano, H., Huppi, K., Heinrich, G. & Tonegawa, S. (1979) *Nature (London)* **280,** 288–294.
9. Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. (1980) *Cell* **19,** 981–992.
10. Tonegawa, S. (1983) *Nature (London)* **302,** 575–581.
11. Givol, D., Zakut, R., Effron, K., Rechavi, G., Ram, D. & Cohen, J. B. (1981) *Nature (London)* **292,** 426–430.
12. Ichihara, Y., Matsuoka, H. & Kurosawa, Y. (1988) *EMBO J.* **7,** 4141–4150.
13. Parslow, T. G., Blair, D. L., Murphy, W. J. & Granner, D. K. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 2650–2654.
14. Falkner, F. G., Mocikat, R. & Zachau, H. G. (1986) *Nucleic Acids Res.* **14,** 8819–8827.