Genet. Res., Camb. (1990), **56**, *pp.* 115–120 *With 5 text-figures Printed in Great Britain*

Grammatical analysis of DNA sequences provides a rationale for the regulatory control of an entire chromosome

SUSUMU OHNO

Beckman Research Institute of the City of Hope, 1450 E. Duarte Road, Duarte, California 91010-0269

(*Received 14 January 1990 and in revised form 3 May 1990*)

Summary

Regardless of their origins, functions, and base compositions, all DNAs are scriptures written following the same grammatical rule. At the level of syllables, two, CG and TA are seldom used, while three, TG, CT and CA are utilized with abundance. Accordingly, at the level of three-letter words, two complementary base trimers, CTG and CAG, invariably enjoy frequent usage. Inasmuch as two of the three frequently used syllables, TG and CA are complementary to each other, while two seldom used syllables, CG and TA, are both palindromes, two complementary strands of DNA are inherently symmetrical with each other. Consequently, palindromic sequences as favourite targets of DNA-binding proteins occur at unsuspectedly high frequencies, if they contain TG and CA or CTG and CAG. Nevertheless, there are grammatical rules operating among these high frequency palindromes as well; e.g. the palindromic tetramer TGCA occurs nearly two times more often than its reciprocal; CATG. Thus, DNA-binding proteins are provided with a wealth of abundant targets whose densities are influenced by a regional difference in GC/AT ratios to variable degrees. One palindromic heptamer CAGNCTG is an ideal target of one DNA-binding protein engaged in chromosome packaging and in generation of banding patterns. This heptamer occurs once every 1000 bases in moderately GC-rich sequences, while its incidence is reduced to once every 3000 bases in extremely AT-rich sequences. The above must be the very reason that a solitary human X-chromosome DNA coated with mouse DNA-binding proteins in mouse-man somatic hybrids still maintains the original banding pattern and that the inactive X remains inactive, while the active X remains active.

1. Introduction

Starting this essay, it suddenly dawned on me that I have had the privilege of knowing Mary F. Lyon for thirty years. Over this long period of acquaintance, her ingeniously simple insights into all the problems she dealt with impressed me time and time again. I do hope she will be amused by what is said in this essay.

The regulatory control involving the entire chromosome is seen not only in mammals with regard to one or the other X of female somatic cells but also in insects and other invertebrates with regard to paternally derived chromosomes. At first glance, this widely spread phenomenon does not appear strange, until the following is considered.

Man-mouse somatic hybrids tend to eliminate human chromosomes in steps; a hybrid cell line which has retained only one human chromosome finally emerging. DNA of such a solitary human chromosome must be associating with DNA binding proteins mostly of the mouse origin. Yet, this chromosome continues to exhibit the human specific banding pattern impervious of the circumstance. Since coding regions stringently conserved by natural selection account for only a few percent of the total DNA, base sequences must be very different even between homologous chromosomes of man and the mouse; e.g. the X chromosome. Yet, the association between mouse DNA-binding proteins and human DNA produces the human specific chromosome banding pattern. Furthermore, the inactive human X remains inactive in such man-mouse hybrids, while the active X remains active. Thus, it would appear that a subset of DNA binding proteins responsible for the maintenance of the inactive state of the X-chromosome also knows no species barrier, in spite of the fact that over all base sequence differences between the human X and the mouse X must be very great indeed. Thus, we are faced with an intriguing paradox of two very long stretches of DNA of the order of 10^8 base pairs that are very different in sequences appearing the same to a set and a subset of DNA binding proteins.

The answer to the above noted paradox is provided below by grammatical analysis of DNA sequences. All DNA base sequences regardless of their origin, or function, are scriptures written following the universal grammatical rule.

2. The universal rule of CG/TA-deficiency and TG/CT/CA-excess

The first rule of all languages is a drastic reduction of the potentially enormous information content to a manageable size. Of the twenty-six letters in the alphabet, the most frequently used letter in English is 'e', as pointed out by the creator of mystery novels; Edgar Allan Poe. In contrast, letters such as 'q', 'x' and 'z' are seldom used. This part of the rule for languages does not apply to DNA, for frequencies of four letters, A, G, T and C, vary depending upon base compositions of DNA. 26 letters in the alphabet yield 676 (26^2) potential syllables. Syllables, here, are defined as two letter combinations. However, certain combinations of consonants such as 'lw', 'wl' or 'ww' are seldom, if ever, used. While words usually begin with syllables that are combinations of vowels and consonants, or vice versa, reciprocal combinations of such syllables are, as a rule, not used in equal frequency. The *Webster's New Collegiate Dictionary* contains nearly 600 words beginning with the syllable 'ho', for example. In sharp contrast, aside from the usage of 'oh!' as an exclamation, the dictionary contains only two words starting with a mirror image of 'ho'; 'ohm' and 'ohmage', both in honour of the German electrician, G. S. Ohm. Were it not for this particular individual, there would not have been a single word of English starting with the syllable 'oh'. On a more modest scale, the same can be said of words starting with 'pa' (about 1000 words) versus those beginning with 'ap' (nearly 300 words). In differential usages of syllables, especially mirror images, one notices, for the first time, the similarity between languages and construction of DNA base sequences. Since four letters, A, G, T and C, can give rise to only 16 syllables (base dimers), all are used, but at very different frequencies. This is illustrated on Figure 1, using the very large human serum albumin gene as one of the two examples. The contiguous sequence is comprised of 19002 bases, its coding regions being represented by only 1827 bases (Minghetti *et al.* 1986). Overall, it is a very AT-rich sequence (65%), although the coding regions are not as AT-rich (57%). The other

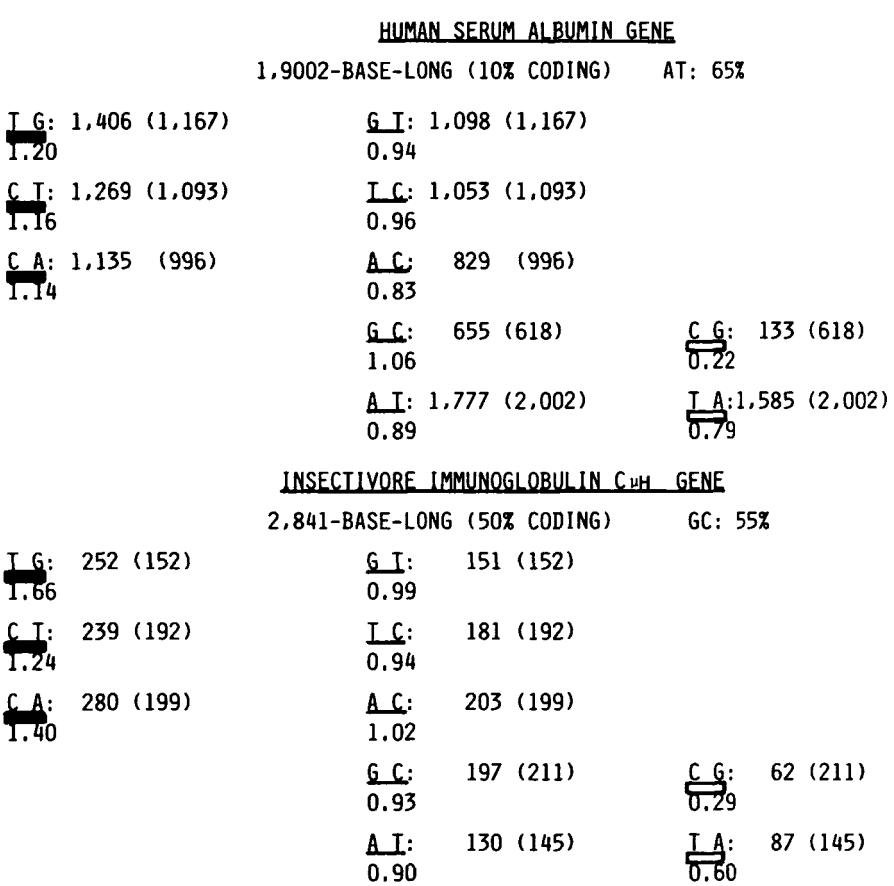

Fig. 1. The universal rule of CG/TA-deficiency and CT/TG/CA-excess, as applied to the human serum albumin gene and insectivore Ig$C\mu_H$ gene is shown with regard to 5 pairs of mirror image base dimers. Each dimer is accompanied by its observed number as well as its expected number in parentheses. Its observed/expected ratio is shown immediately below. Always over-represented dimers are underlined by thick solid bars, while consistently deficient dimers are underlined by thick open bars.

https://doi.org/10.1017/S0016672300035187 Published online by Cambridge University Press

example used is the 2841 base-long gene sequence for immunoglobulin μ-heavy chain constant region of the insectivore (Ishiguro *et al.* 1989). Nearly half of the sequences are represented by the coding regions. This gene is GC-rich (55%) overall; the coding region particularly; GC: 65%. Thus, the human serum albumin gene and the insectivore Ig $C_{\mu H}$ gene provide a nice contrast with regard to base compositions as well as ratios between coding and noncoding regions.

Observing Fig. 1, be aware that in spite of these marked differences noted above, DNA sequences of these two genes obey the same grammatical rule as to differential uses of 16 syllables (base dimers). Three, TG, CT, CA, are over-utilized, whereas two, CG and TA, are markedly under-utilized. Their mirror image dimers, on the other hand, are inconsistently utilized, usually hovering around their expected rates computed from base compositions of DNA sequences. Inasmuch as 10 dimers are covered in Fig. 1, only 6 dimers remain, four of which are homodimers; AA, GG, TT and CC. This, then, is the universal rule of CG/TA deficiency-TG/CT/CA excess originally proposed as the rule for coding sequence construction (Ohno, 1988), but later was found to cover all DNA sequences including non-coding regions (Yomo & Ohno, 1989).

3. CTG as the equivalent of the word 'the'

A priori, one would think that shorter words are more frequently utilized than longer words in all languages. Indeed, there are a number of words in monosyllables such as the preposition 'to' and 'do' and 'go' as verbs in English. Yet, the most frequently used word in English is 'the', as pointed out again by Edgar Allan Poe in his '*The Gold-bug*'. In addition, this three letter combination is included as parts of numerous words such as 'theorem', 'their', 'them', 'atheist', 'gather', 'feather', to the extent that if the use of this three letter combination is precluded, one is hard pressed to compose a sensible paragraph in English. DNA base sequences also have the base trimer equivalent of 'the'. Inasmuch as the base trimer CTG is a composite of two always excessive dimers, CT and TG, it is one of the most consistently overabundant base trimers in all DNA. Not far behind in terms of overabundance, is its complementary base trimer CAG. In extremely

INSECTIVORE IMMUNOGLOBULIN $C_{\mu H}$ GENE
2,841-BASE-LONG (50% CODING) CG: 55%

1)	C T G 2.23	105 (47)	4)	C A G 1.76	86 (49)	30B)	G A A 1.08	40 (37)	39)	T T C 0.83	35 (42)
2)	C A C 1.59	97 (61)	11)	G T G 1.81	67 (37)	32B)	G T C 0.83	39 (47)	34A)	G A C 0.78	38 (49)
•3)	C C T 1.61	95 (59)	•26A)	A G G 1.18	46 (39)	•34B)	A A C 0.83	38 (46)	•51A)	G T T 0.65	22 (34)
•5)	C C A 1.39	85 (61)	•22A)	T G G 1.49	55 (37)	34C)	C T T 0.90	38 (42)	38)	A A G 0.97	36 (37)
6)	C C C 1.04	84 (81)	19)	G G G 1.41	58 (41)	37)	C A A 0.80	37 (46)	41A)	T T G 0.88	30 (34)
7)	T G C 1.60	79 (47)	14A)	G C A 1.27	62 (49)	41B)	T A C 0.68	30 (44)	50A)	G T A 0.66	23 (35)
•8A)	G C C 1.20	78 (65)	•26B)	G G C 0.90	46 (51)	43A)	A T C 0.66	29 (44)	48)	G A T 0.71	25 (35)
8B)	C T C 1.32	78 (59)	14B)	G A G 1.59	62 (39)	43B)	A T A 0.88	29 (33)	46A)	T A T 0.81	26 (32)
10)	A C A 1.65	76 (46)	13)	T G T 1.91	65 (34)	45)	A A A 0.80	28 (35)	46B)	T T T 0.84	26 (31)
•12)	A C C 1.08	66 (61)	•40)	G G T 0.86	31 (37)	49)	T A A 0.73	24 (33)	55A)	T T A 0.53	17 (32)
16)	C A T 1.36	60 (44)	25)	A T G 1.42	50 (35)	50)	C C G 0.35	23 (65)	59A)	C G G 0.30	16 (56)
17A)	A G C 1.20	59 (49)	28)	G C T 0.96	45 (47)	51B)	A T T 0.71	22 (31)	53)	A A T 0.58	19 (33)
17B)	T C C 1.00	59 (59)	30A)	G G A 1.02	40 (39)	54)	C T A 0.41	18 (44)	55B)	T A G 0.49	17 (35)
20A)	A G A 1.54	57 (37)	22B)	T C T 1.31	55 (42)	•55C)	C G A 0.35	17 (49)	•64)	T C G 0.21	10 (47)
20B)	T C A 1.30	57 (44)	24)	T G A 1.51	53 (35)	55D)	A C G 0.35	17 (49)	59B)	C G T 0.34	16 (47)
29)	A C T 1.00	44 (44)	32A)	A G T 1.11	39 (35)	62)	C G C 0.20	13 (65)	63)	G C G 0.21	12 (56)

Fig. 2. 64 base trimers contained in the insectivore Ig$C\mu_H$ are labelled 1-to-64 in the order of their abundances. When two or more trimers are equal in abundance they are given the same number (e.g. no. 17a and no. 17b), the trimer next in the order of abundance being labelled no. 19. 64 are arranged as 32 pairs of complementary trimers, no. 1 occupying the top of the left column, while its complementary trimer ranking no. 4, occupied the top of the 2nd from the left column. At the other extreme, no. 62 trimer occupied the bottom of the 2nd from right column, while its complementary trimer ranking no. 63 is found at the bottom of the right column. Their observed and expected (in parentheses) numbers, as well as observed/expected ratios are shown. Asterisks are explained in the text.

https://doi.org/10.1017/S0016672300035187 Published online by Cambridge University Press

AT-rich sequences, however, these two are overshadowed by such base trimers as AAA, AAT, and TTT in terms of absolute numbers, although not in degrees of overabundance. Conversely, they are dominated by such base trimers as GGG, GGC and CCC in extremely GC-rich sequences.

In DNA of rather balanced base compositions (i.e. GC: $50 \pm 5\%$), either CTG or CAG enjoy the numerical superiority as well. Listed in Fig. 2 are observed numbers, expected numbers, as well as observed/expected ratios of all 64 base trimers contained in the 2841-base-long gene sequence for the insectivore immunoglobulin μ-heavy chain constant region with 55% GC (Ishiguro *et al.* 1989). These 64 base trimers are ranked 1 to 64 from the most numerous to the least and arranged as 32 pairs of complementary trimers.

It should be noted that CTG, which is the most over-represented of the 64 base trimers (2·23 × the expected) is also the most numerous (102) and that its complementary base trimer CAG is not far behind, ranking 4th. As a rule, those base trimers containing TG, CA or CT belong in the top half, while those containing TA or CG belong in the bottom half. Although slightly more than half of the base trimers are derived from noncoding regions, trimer frequencies shown in Fig. 5 faithfully reflected codon usages. Thus, of 6 codons for leucine, the most frequently used is the most numerous CTG, while 55th TTA and 54th CTA are the least utilized Leu codons. Similarly, of 4 potential codons for Proline, CCT, CCA and CCC ranked 3rd, 5th and 6th, while CCG ranked 50th. Accordingly, only 4 of the 37 Prolines as 4th most numerous residue after Ser, Thr, Leu are encoded by CCG.

Because two of the three invariably over-represented base dimers, TG and CA, are complementary to each other and because both of the two constantly deficient dimers, TA and CG, are palindromes, 32 pairs of complementary base trimers demonstrate remarkable symmetry. Excluding 5 pairs marked by asterisks, Fig. 2 shows that differences in numbers of complementary base trimers of 27 pairs are less than 30%. In the case of 7 pairs, two complementary base trimers are nearly equal in numbers; e.g. 57 AGA versus 55 TCT and 17 ACG versus 16 CGT. Thus, the abundance of palindromic base oligomers is the rule rather than exception of all DNA, as remarkable symmetry is invariably maintained between two complementary strands that comprise DNA (Yomo & Ohno, 1989).

4. Sequential orders of words and sequential orders of four bases in forming palindromic tetramers

In all languages, reciprocal combinations of two words such as 'we do' and 'do we' bring forth different meanings, and in certain instances, one remains sensible, while the other becomes nonsensical; e.g. 'we go' and 'go we'. Similarly in DNA, the tetramer CCTG is very common whereas TGCC is found far less frequently. Here, however, we shall limit our discussion to palindromic sequences. Since the *lac* operator base sequence of Jacob and Monod, the preference of DNA-binding proteins to recognize palindromic base oligomers is well established. This is no surprise, for due to its double-strandedness, only a palindromic piece of DNA would appear as a single entity to proteins. Various palindromes of the same base compositions occur at very different rates. In languages with 676 (26^2) potential syllables such as English, palindromic words are not very common and rarer still are palindromic sentences; e.g. 'Able was I, ere I saw Elba.' Nevertheless, 'deed' is a palindromic word, whereas its rearranged palindrome 'edde' is not a word. When 'edde' is encountered, it can only represent a junction between two adjoining words; e.g. 'imposed denial'.

Keeping the above in mind, we shall now examine

FREQUENCIES OF 8 BALANCED PALINDROMIC TETRAMERS IN TWO GENES

HUAL: 1,9002-BASE-LONG HUMAN SERUM ALBUMIN GENE AT:65%

INIG: 2,841-BASE-LONG INSECTIVORE IgCμH GENE GC:55%

( ): NUMBER/10,000 BASES

	HUAL	INIG		HUAL	INIG
1) TGCA:	82 (43.1)	22 (77.4)	2,3) CATG:	52 (26.8)	9 (31.7)
3,2) AGCT	44 (23.1)	15 (52.7)	4) CTAG:	26 (13.7)	4 (14.1)
5,7) GATC	22 (11.6)	1 (3.5)	7,8) TCGA:	8 (4.2)	0 (----)
6) GTAC	15 (7.9)	3 (10.6)	8,5) ACGT	7 (3.7)	5 (18.6)

Fig. 3. Using the same two genes analysed in Fig. 1, 8 tetrameric palindromes containing one each of 4 bases are ranked in the order of their abundance, 1–8. When the same tetramer ranked differently in two genes, both rankings are shown. They are arranged as 4 reciprocal pairs; e.g. TGCA and CATG as a pair. Their observed numbers are accompanied by numbers (in parentheses) adjusted to per 10000 bases. The TGCA tetramer is underlined with a thick solid bar; whereas palindromic tetramers containing CG are underlined with thick open bars. The TA portion of the palindromic tetramers is underlined with thin open bars. HuA1: 19002-base-line human serum albumin gene. InIg: 2841-base-long insectivore immunoglobulin μ-heavy chain constant region gene.

https://doi.org/10.1017/S0016672300035187 Published online by Cambridge University Press

palindromic base tetramers containing one each of four bases. There can be 4 pairs; TGCA and CATG, GTAC and ACGT, AGCT and CTAG as well as GATC and TCGA. Their frequencies are shown in Fig. 3 with regard to both the 19002-base-long human albumin gene which is 65% AT (Minghetti *et al.* 1986) and the 2841-base-long insectivore immunoglobulin μ-heavy chain constant region gene which is 55% GC (Ishiguro *et al.* 1989). It should be noted that in spite of a 25% difference in the GC contents of two genes, the orders of abundance among 8 tetrameric palindromes are about the same. Because of the complementary between two of the three invariably excessive dimers, either TGCA or CATG are expected to be the most numerous of the 8 tetramers. Yet, there is a grammatical rule operating in this pair as well. For TGCA is nearly twice or more abundant than CATG in all genes; TGCA corresponding to 'we do' and CATG to 'do we'. Due to the severe CG-deficiency in all genes, the least abundant are TCGA, or ACGT; GTAC and CTAG suffering the similar fate because of TA-deficiency. These four probably correspond to a nonsensical palindrome 'edde', while GATC and AGCT correspond to a palindromic word, 'deed'.

5. Palindromic hexamers, heptamers and octamers as recognition targets of DNA-binding proteins

Palindromic tetramers are too short a target for DNA-binding proteins, the preference by proteins being hexamers to octamers. Of 4 different palindromic hexamers centred by TGCA, two are AT-rich (2 A's and 2 T's), while the other two are GC-rich (2 G's and 2 C's). Since the serum albumin gene is very AT-rich, TTGCAA and ATGCAT hexamers are found in abundance there (Fig. 4). Although slightly GC-rich, the IgCμH gene contained none of the two. This absence, however, seems to be an accident derived from the shortness of its length. As computed from two overlapping pentamers, 5 ATGCA and 4 TGCAT, this Ig gene is expected to have contained 5·2 ATGCAT hexamers per unit length of 10000 bases which is more than 2·1 per 10000 bases for the serum albumin gene. At any rate, our experience with a large number of genes from diverse organisms indicated these two AT-rich palindromic hexamers to be freelancers that obey no universal rule as to their incidences. They probably correspond to such very specialized words as 'bridge' and 'swells' which occur quite frequently only in writings pertaining to the subject of water. As to consistency, the most reliable palindromic hexamer is CTGCAG. Since this hexamer is 57% GC, it is found more frequently in GC-rich sequences (Fig. 4). Nevertheless, at least one copy of it is bound to be found per unit length of 10000 bases even in very AT-rich sequences (Fig. 4). Interestingly, CAGCTG is not as common as CTGCAG, as shown in Fig. 4. This is simply because TGCA, at the centre, is more numerous than AGCT (Fig. 3). Since there are not many commonly used six-letter words, CTGCAG probably corresponds to very common two-word combinations such as 'all the'.

It would thus appear that the CAGCTG hexamer is one of the ideal targets for DNA-binding proteins engaged in chromosome packaging as well as in generation of banding patterns. This hexamer is expected to occur once every 6000 bases, even in a very AT-rich DNA sequence, whereas its incidence increases to once every 700 bases in a moderately GC-rich sequence. Before leaving the subject of palindromic hexamers, the 'TATA' box as the transcription initiation site for RNA polymerase II deserves a notice. The 'TATA' box hexamer is, in a literal sense, a TATATA hexamer. Due to the universal TA-deficiency, this hexamer is expected to be one of the rarest. Yet, TA-deficiency is less severe in noncoding regions than in coding regions (Ohno & Yomo, 1990). Accordingly, the 19002-base-long human serum albumin gene, which consists mainly of non-coding regions (90%) and extremely AT-rich, contained 7 TATATA as depicted at the extreme right

	HuAl	InIg		HuAl	InIg
TGCA:	82 (43.1)	22 (77.4)			
TTGCAA:	10 (5.3)	*0 (1.7)	TATATA	7 (3.6)	1 (3.5)
ATGCAT:	4 (2.1)	*0 (5.2)			
CTGCAG:	3 (1.6)	4 (14.1)			
GTGCAC:	2 (1.1)	2 (7.0)			
AGCT:	44 (23.1)	15 (52.7)			
CAGCTG:	1 (0.5)	2 (7.0)			

Fig. 4. With regard to the same two genes, observed numbers as well as numbers per 10000 bases (in parentheses) of four palindromic hexamers containing TGCA at the centre are shown in the top four rows of the centre column. Shown below them are those of CAGCTG; a reciprocal of the CTGCAG hexamer. At the extreme right, observed numbers, as well as numbers per 10000 bases of the TATATA hexamer are shown. Asterisks are explained in the text.

https://doi.org/10.1017/S0016672300035187 Published online by Cambridge University Press

of Fig. 4. Even the insectivore IG CμH gene, which is only 2841-bases-long, contained one TATATA in its 3′ non-coding region. Thus, the 'TATA' box hexamer is not a unique sequence to be found only in the 5′ non-coding region of each gene. This problem of commonality with regard to the 'TATA' box hexamer is greatly complicated further by innumerable base substitutions that are permissible. Hence, TTAAAT and TTAAAA are but two of the many hexamers exclusively made of A and T that serve as transcription initiation signals for RNA polymerase II. The problem of commonality is by no means confined to the 'TATA' box. It is a problem faced by all regulatory palindromes (Ohno & Yomo, 1990).

More common palindromic heptamers having an odd base at the centre are listed in Fig. 5. Of those, CAGNCTG where 'N' is most often 'T' appears to be an even better candidate than the already noted CTGCAG hexamer as a signal sequence for DNA-binding proteins engaged in chromosome packaging and the generation of banding patterns. This heptamer is more evenly distributed than CTGCAG between AT-rich genes and GC-rich genes; e.g. once every 3000 bases for an extremely AT-rich sequence and once every 1000 bases for a moderately GC-rich sequence. There is a universal grammatical rule operating among these heptamers as well. In all genes, we found CAGNCTG heptamers to be more numerous than CTGNCAG. It would be recalled that the rule was reverse with regard to palindromic

	HUAL	INIG
CAGNCTG:	7 (3.7)	3 (10.6)
CTGNCAG:	3 (1.6)	2 (7.0)
CACTGTG:	2 (1.1)	1 (3.5)
CACTGTT:	4 (2.1)	1 (3.5)
CTCTGTG:	2 (1.1)	1 (3.5)
CACATGTG:	2 (1.1)	1 (3.5)
CACTGGTG:	1 (0.5)	1 (3.5)
TGCATGCA:	1 (0.5)	1 (3.5)

Fig. 5. Shown in the first 3 rows are incidences in the same two genes of three palindromic heptamers having an odd base at the centre. The third is boxed in because of its significance as a signal sequence for the long distance recombination between various components of immunoglobulin genes (see the text). Shown immediately below this heptamer are incidences of two of its single base substituted versions. Shown in the 6th and 8th rows are incidences of two palindromic octamers made exclusively of CA and TG. Sandwiched between these two octameric palindromes is CACTGGTG. This octamer is a two-base substituted version of the octamer in the 6th row. At the same time, it is also a single-base inserted version of the boxed-in heptamer CACTGTG.

hexamers; CTGCAG always outnumbering CAGCTG. Of the other heptamers related to the above, the most noteworthy is CACNGTG, for this is the heptamer used for the long distance recombination between various components of immunoglobulin as well as T-cell receptor genes. This heptamer where 'N' is 'A' is more often found at the coding region 3′ end of V_L, V_H, and D_H segments, whereas the same heptamer where 'N' is 'T' is more often found at coding sequence 5′ ends of J_L, and D_H as well as J_H (Sakano *et al.* 1979; Early *et al.* 1980). Here again, we are faced with the problem of commonality, for CACNGTG is by no means a rare heptameric palindrome as shown in Fig. 5. The problem is further compounded by a number of permissible, single base substitutions; equally common CACTGTT and CTCTGTG (shown in Fig. 5) also serving as heptameric signals for the long distance recombination (Ohno & Yomo, 1990). It is probable that a key component of the recombinase complex that mediates this long-distance recombination arose from the universal DNA-binding protein that engaged in chromosome packaging by recognizing the CAGNCTG heptamer. Admittedly, CACNGTG is not as common as CAGNCTG. Nevertheless, it is common enough to create the problem of inadvertent, therefore, illegitimate recombinations (Ohno & Yomo, 1990). Interestingly, an insertion of A to this CACTGTG heptamer at the position between CAC and TGTG created one of the more common palindromic octamers; CACATGTG (Fig. 5).

References

Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. (1980). An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: V_H, D, and J_H. *Cell* **19** 981–992.

Ishiguro, H., Ichihara, Y., Namikawa, T., Nagatsu, T. & Kurosawa, Y. (1989). Nucleotide sequences of *Suncus Murinus* immunoglobulin μ gene and comparisons with mouse and human μ genes. *FEBS Letters* **247**, 317–322.

Minghetti, P. P., Ruffner, D. E., Kuang, W.-J., Dennison, O. E., Hawkins, J. W., Beattie, W. G. & Dugaiczyk, A. (1986). Molecular structure of the human albumin gene is revealed by nucleotide sequence within q11-22 of chromosome 4. *Journal of Biological Chemistry* **261**, 6747–6757.

Ohno, S. (1988). Universal rule for coding sequence construction: TA/CG deficiency-TG/CT excess. *Proceedings of the National Academy of Sciences USA* **85**, 9630–9634.

Ohno, S. & Yomo, T. (1990). Various regulatory sequences are deprived of their uniqueness by the universal rule of TA/CG-deficiency and TG/CT excess. *Proceedings of the National Academy of Sciences USA* **87**, 1218–1222.

Sakano, H., Huppi, K., Heinrich, G. & Tonegawa, S. (1979). Sequences at somatic recombination sites of immunoglobulin light-chain genes. *Nature* **280**, 288–294.

Yomo, T. & Ohno, S. (1989). Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. *Proceedings of the National Academy of Sciences USA* **86**, 8452–8456.