J Mol Evol (1996) 43:125–134

JOURNAL OF MOLECULAR EVOLUTION

Sequence Simplicity and Evolution of the 3′ Untranslated Region of the Histone H1° Gene

Imma Ponte,[1] Claudio Monsalves,[2] Miguel Cabañas,[1] Pedro Martínez,[3] Pedro Suau[1]

[1] Departamento de Bioquímica i Biología Molecular, Facultad de Ciencias, Universidad Autónoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
[2] Departamento de Biología Molecular, Facultad de Ciencias Biológicas, Universidad de Concepción, Chile
[3] Division of Biology, California Institute of Technology, Pasadena, California 91125

Received: 14 September 1995 / Accepted: 25 January 1996

Abstract. The H1° gene has a long 3′ untranslated region (3′UTR) of 1,125 nucleotides in the rat and 1,310 in humans. Analysis of the sequences shows that they have features of simple DNA that suggest involvement of replication slippage in their evolution. These features include the length imbalance between the rat and human sequences; the abundance of single-base repeats, two-base runs and other simple motifs clustered along the sequence; and the presence of single-base repeat length polymorphisms in the rat and mouse sequences. Pairwise comparisons show numerous short insertions/deletions, often flanked by direct repeats. In addition, a proportion of short insertions/deletions results from length differences in conserved single-base repeats. Quantification of the sequence simplicity shows that simple sequences have been more actively incorporated in the human lineage than in the rodent lineage. The combination of insertions/deletions and nucleotide substitutions along the sequence gives rise to three main regions of homology: a highly variable central region flanked by more conserved regions nearest the coding region and the polyA addition site.

Key words: Histone H1° — 3′ untranslated region — Slippage — Sequence simplicity — Evolution

Correspondence to: P. Suau

Introduction

Histone H1° is a chromosomal protein belonging to the H1 histone class. In addition to H1°, five somatic subtypes, H1a, b, c, d, and e (Lennox and Cohen 1983), and a germinal-line-specific subtype, H1t (Seyedin and Kistler 1979), have been identified in mammals. It is currently accepted that H1 proteins could be involved in the regulation of transcription through the modulation of chromatin higher-order structure. Preferential binding to scaffold-associated regions and participation in nucleosome positioning have also been proposed as other possible mechanisms by which H1 could contribute to transcriptional regulation (Izaurralde et al. 1989; Meersseman et al. 1991; Zlatanova and Doenecke 1994). Several features of the expression, genomic organization, and transcript processing confer H1° a distinct position among histone subtypes. H1° is the only histone subtype that is transcriptionally activated at terminal differentiation (Ponte et al. 1994). H1° expression is also modulated by external signals in specific populations of mature cells (Gjerset et al. 1982; García-Segura et al. 1993; Lafarga et al. 1995). Histone genes are organized in clusters containing one gene of each of the five histone classes (Hentschell and Birnstiel 1981). The clusters are reiterated a variable number of times depending on the species (Heintz et al. 1981). In contrast, H1° is a single-copy gene and it is not clustered with core histones. In humans, the H1° gene maps to chromosome 22, while the other H1 genes, together with core histone genes, are located on chromosome 6 (Albig et al. 1993). Here we

analyze the 3′ untranslated region (3′UTR) of the H1° gene from the rat, the mouse. and humans (Doenecke and Tönjes 1986; Alonso et al. 1988; Castiglia et al. 1993; Brocard et al. 1994; Martinez et al. 1995). The 3′UTR of eukaryotic transcripts can have multiple effects on gene expression, including the control of mRNA stability, the regulation of translation efficiency, and the cytoplasmic localization of mRNA (Decker and Parker 1995). The 3′UTR is usually very short. Core histones and all the H1 subtypes except H1° have 3′UTRs of a few hundred nucleotides. The H1° gene has an unusually long 3′UTR, of 1,125 nucleotides in the rat and 1,310 nucleotides in humans (Doenecke and Tönjes 1986). The 3′UTR of H1° thus offers the opportunity to study the constraints, evolutionary stability, and sources of genetic variation involved in the evolution of long 3′UTRs. We show that the 3′UTR of H1° can be considered as an example of the simple sequences that are frequently found in eukaryotic genomes. These sequences are characterized by a clustered distribution of short nucleotide motifs (1–5 bp), which suggests involvement of slippage in its evolution (Tautz and Schlötterer 1994; Dover 1995). Slippage has been shown to act on both coding (Eickbush and Burke 1986; Djian and Green 1989; Treiter et al. 1989; Paulsson et al. 1990; Costa et al. 1991) and noncoding sequences (Hancock and Dover 1988; Hoelzel et al. 1991). Some features of the 3′UTR of the H1° gene, including abundant irregular direct repeats, a high incidence of short insertions/deletions (≤20 bp), and a length imbalance between the rat and human sequences, suggest that, in addition to nucleotide substitutions, the errors in DNA replication of the kind of replication slippage have had an important role in its evolution.

Materials and Methods

Cloning of a Rat H1° cDNA From PC12 Cells. A cDNA library in the vector λSWAJ (Palazzolo and Meyerowitz 1987) was prepared using polyA+ RNA from rat PC12 cells, differentiated with nerve growth factor. Positive clones were selected using a mouse H1° cDNA probe and subcloned in Bluescript (Stratagene). The longest clone had 1.8 kb. The entire sequence of both strands was obtained by the dideoxynucleotide chain-termination method using a *Pharmacia* sequencing kit.

Sequence Analysis. Sequence analysis was performed with the software package GCG (Genetic Computer Group Inc., University of Wisconsin, version 8) on a VAX computer. Sequences were aligned with the help of a series of dot matrix comparisons for the definition of homology blocks and with the program BESTFIT, which uses the local homology algorithm of Smith and Waterman (1981). Sequences were obtained from the EMBL database. The sequences analyzed are listed in the caption of Fig. 1.

Frequency Distribution of Runs of a Single Base. Frequency distributions of single-base repeats of different length (2–12 nucleotides) were calculated for the rat and human sequences. Theoretical frequency distributions were calculated by two methods: (1) Using the nucleotide composition according to the expression:

$$\sum^{i} f(N_n^i) = \sum^{i} (fN^i)^n$$

where, $f(N_n^i)$ is the theoretical frequency of a single-base repeat of length n; N^i = (A,T,C,G) and fN^i = the nucleotide frequency. (2) Using the dinucleotide composition of the test sequences. This calculation is based on the general expression of Markov processes:

$$f(N_1,N_2,N_3,\ldots N_n) = f(N_1)f(N_2|N_1)f(N_3|N_2)\ldots f(N_n|N_{n-1})$$

Conditional probabilities were calculated using the dinucleotide composition of the test sequences with the expression:

$$\sum^{i} f(N_n^i) = \sum^{i} (fN^i)\left(\frac{f(N^iN^i)}{f(N^iN^j)}\right)^{n-1}$$

where, $f(N_n^i)$ is the theoretical frequency of single-base repeats of length n; N^i = (A,T,C,G); fN^i, the nucleotide frequency; $f(N^iN^i)$, the dinucleotide frequency; $f(N^iN^j)$, the observed frequency of all possible dinucleotides with N^i as the first nucleotide and N^j = (A,T,C,G).

Sequence Simplicity. The clustering of tri- and tetranucleotide motifs either in short arrays or interspersed along the sequences was estimated with the program SIMPLE34 (Hancock and Armstrong 1994). The program was obtained by anonymous ftp from life.anu.edu.au. SIMPLE34 is a modified version of SIMPLE originally described by Tautz et al. (1986). The algorithm assigns a simplicity score (SS) to individual nucleotides, which is a measure of the abundance of tri- and tetranucleotide motifs starting to the right of each nucleotide inside a window of ±32 nucleotides. Overall simplicity factors (SF) are calculated by summing all scores and dividing the sum by the number of nucleotides. Relative simplicity factors (RSF) are obtained by dividing the overall simplicity factors of the test sequence by the mean SF for ten random sequences of the same length and nucleotide and dinucleotide composition as the test sequence. The RSF of sequences showing the same amount of motif clustering as random sequences should be close to 1.000 and significantly greater for "simple" sequences. The standard deviation of the SFs of the ten random sequences allows the analysis of the statistical significance of the RSF. Three confidence limits are returned by program: 99.7% ($P < 0.003$), 99.0% ($P < 0.01$), 95.0% ($P < 0.05$). In addition, the program identifies simplicity scores associated with individual motifs that are significantly higher (90%) than would be expected by chance.

Results

Polyadenylation Signals

The mRNA of histone H1° is polydenylated. The polyadenylation signal, AAUAAA, was found ten nucleotides from the polyA tail in both the rat and humans. Another AAUAAA motif was found at 106 nucleotides from the polyA tail in the rat and at 126 nucleotides in humans. Both elements are surrounded by a similar context of short direct repeats, which is indicated by arrows in Fig. 1. The repeats keep the same order about both motifs, but they are spaced by sequences of variable length. The direct repeats are highly conserved between rat and humans. Such a context of conserved direct repeats suggests that the presence of two signals could

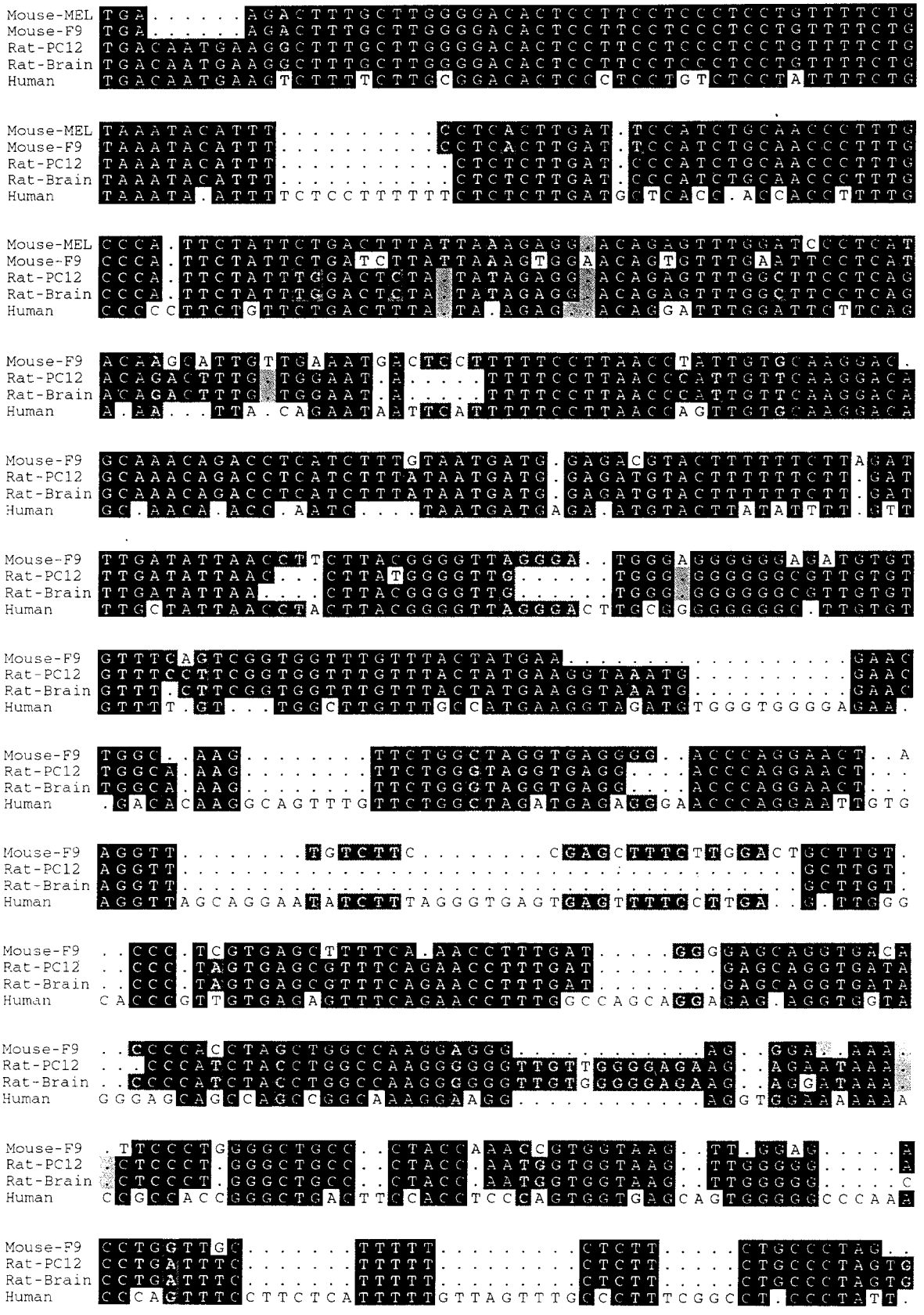

Fig. 1. Alignment of the 3′UTR of histone H1° from the mouse, rat, and humans. Homologous positions are *boxed in black or dark grey* and gaps are indicated by *points*. The polyadenylation signals are indicated by *empty boxes*. The *arrows* numbered from 1 to 5 indicate the conserved context of short direct repeats surrounding the polyadenylation motifs. Gaps due to the different length of conserved single-base repeats are in *light grey*. The accession numbers in the EMBL nucleotide sequence database are: mouse MEL (murine erythroleukemia cells), X72894; mouse F9 (murine teratocarcinoma cells), X13171; rat PC12 (pheochromocytoma cells), X72624; rat brain (cDNA), X70685; human (genomic), X03473.

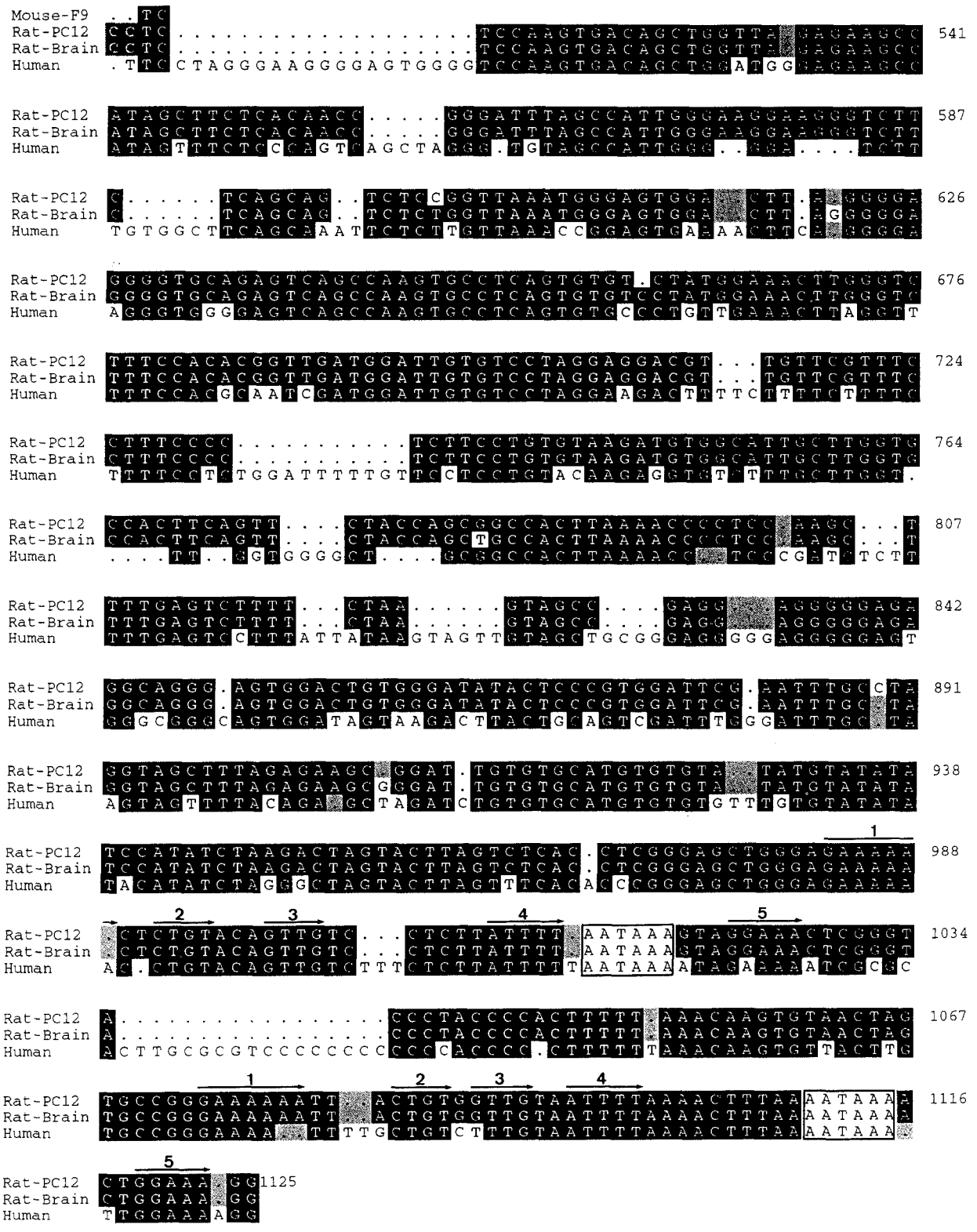

Fig. 1. Continued.

result from a tandem duplication of the downstream sequences of the 3′UTR. It remains to be seen whether the upstream element can be used as an alternative polyadenylation signal in the rat and humans. In the mouse, a functional noncanonical polyadenylation signal has been described at 101 nucleotides downstream of the stop codon (Brocard et al. 1994). This additional signal is not present in the rat and humans.

Comparison of the Nucleotide Sequences

We have aligned the complete 3′UTRs of the H1° gene from rat brain (Castiglia et al. 1993) and rat PC12 cells

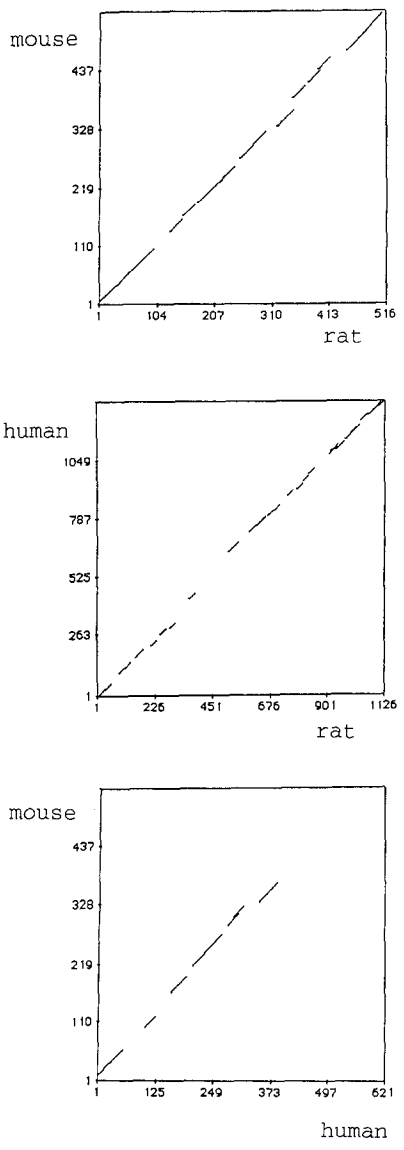

Fig. 2. Dot matrix comparisons of the 3′UTR of histone H1° from the mouse F9 cells, rat PC12, and humans. Match stringency was 17 out of 24 analyzed.

(Martínez et al. 1995) with the human sequence (Doenecke and Tönjes 1986). A partial sequence of 519 nucleotides from mouse teratocarcinoma cells (Alonso et al. 1988) and a sequence of 121 nucleotides from murine erythroleukemia cells (Brocard et al. 1994), resulting from the usage of the mouse-specific polyadenylation signal, have also been included (Fig. 1). The alignment is based on the homology blocks displayed on dot matrix comparisons like that shown in Fig. 2, and was achieved with the help of the BESTFIT program of the GCG package. Pairwise comparisons of the mouse, rat, and human sequences show that the 3′UTR evolves by insertion/deletion as well as by simple nucleotide substitution. Insertions/deletions manifest in dot matrix comparisons by the horizontal and vertical shifts of the homology blocks that define the diagonal (Fig. 2). The 3′UTRs from the rat and humans show an important length imbalance of 185 nucleotides, the rat sequence having

```
mouse    CGGGGTTAGGGATGGGAG
rat      TGGGGTTG....TGGG.G

mouse    GAA........GAA
rat      GAAGGTAAATGGAA

mouse    GGAGGG............AGGGA
rat      GGGGGGTTGTTGGGGAGAAGAGA

rat      TCTT.....CTGCCCT
human    CCTTTCGGCCT.CCCT

rat      CTTC......TCA
human    CTTTGTGGCTTCA

rat      TTTCCCC...........TCTTCCT
human    TTTCCTCTGGATTTTTGTTCCTCCT

rat      GTAGCC.......GAGGAGGGGGA
human    GTAGCTGCGGGAGGGGGAGGGGGA

rat      TTGGTGCCACTTCAGTT
human    TTGGT.....TT..GTT

rat      TAA......GTAG
human    TAAGTAGTTGTAG

rat      TTT...........CTCTC
human    TTTTCTCCTTTTTTCTCTC
```

Fig. 3. Examples of insertions/deletions flanked by short direct repeats within 3′UTR of the histone H1° gene. Short direct repeats near the ends of the gaps are *underlined.*

1,125 nucleotides and the human sequence 1,310 nucleotides. The accommodation of that length difference and the abundance of insertions/deletions were the main determinants of the alignment. Examination of the nucleotide sequences surrounding the putative insertion/deletion sites revealed the presence of short (3–7 bp) direct repeats. These repeats may have been involved in the generation of the insertions/deletions through slippage mechanisms. Several examples of short direct repeats near the ends of each insertion/deletion are shown in Figure 3.

In order to estimate homologies, insertions/deletions were considered as single mutational events irrespective of the length. This method of treating gaps is similar to that used by Miyata et al. (1980) in the calculation of the sequence difference of noncoding regions. According to our alignment, the rat and human sequences are 78% identical. The degree of homology is not uniform along the sequences. Three main homology regions were identified on dot matrix comparisons, which were approximately delimited by the positions 1–300, 301–500, and 501–1125 of the rat sequence (Fig. 1). The regions near-

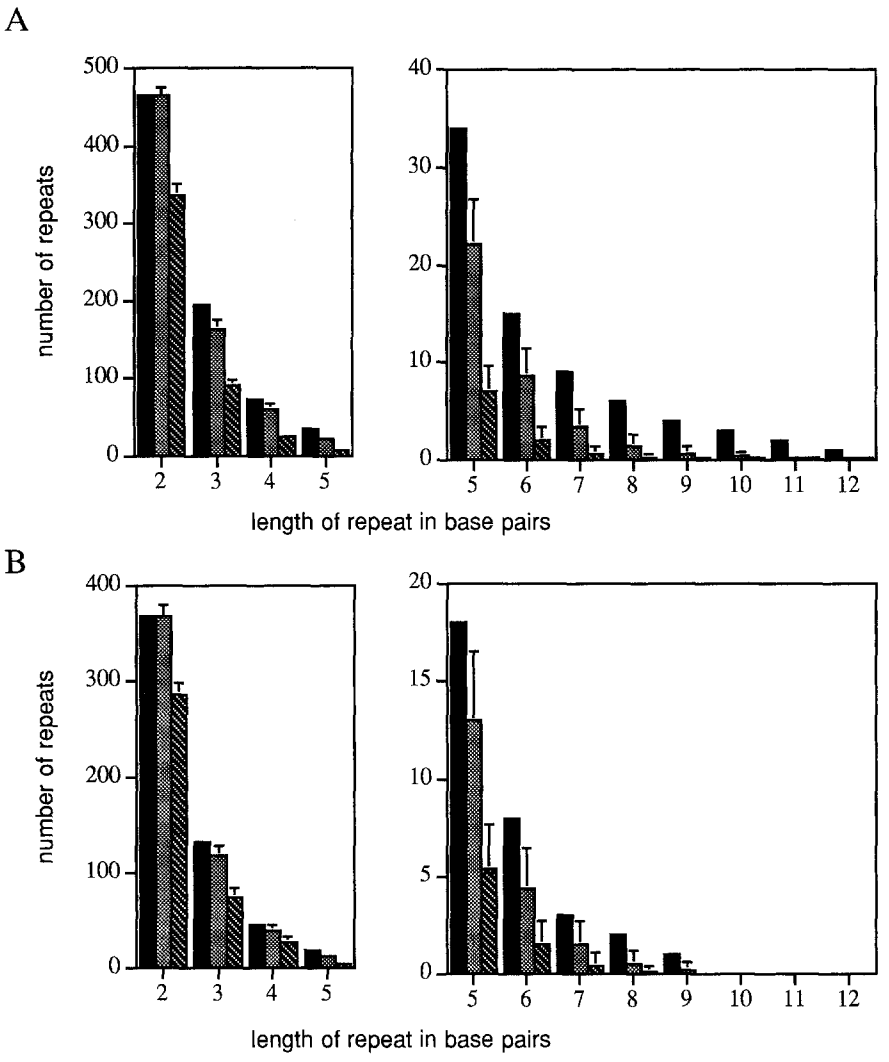

Fig. 4. Frequency distributions of repeats of a single base in the 3′UTR of the histone H1° gene from the rat **(A)** and humans **(B)**. The total number of repeats of length 2 to 12 is indicated by the *bars*. *Solid bars* represent the data for the 3′UTR; *stippled bars* the theoretical frequencies calculated from the dinucleotide composition; and *hatched bars* the theoretical frequencies calculated from the nucleotide composition.

est the stop codon and the polyA addition site have overall homologies of 78% and 80%, respectively, while the middle region is significantly less conserved, with 69% homology. Although the middle region represents less than 20% of the length of the 3′UTR, it absorbs about 60% of the imbalance of 185 nucleotides between the rat and human sequences, as manifested in dot matrices by the vertical displacement of the flanking regions (Fig. 2). The lower homology of the middle region is due to a relatively higher frequency of both insertions/deletions and nucleotide substitutions.

Single-Base Repeats

Single-base repeats are abundant in the H1° 3′UTR. In the human sequence, runs of up to 12 nucleotides are present. It has been reported that introns contain a substantial excess of longer runs of a single base than would be expected on a random basis (Blaisdell 1983; Levinson and Gutman 1987a). The abundance of single-base runs was taken as an argument in favor of the involvement of intrahelical events of the kind of replication slippage in the evolution of intron sequences (Levinson and Gutman 1987a). We have performed a similar analysis on the 3′UTR of the H1° gene. In Fig. 4 the frequencies of runs of a single base in the natural sequences are compared with the theoretical frequencies. The theoretical frequencies were calculated using two procedures. In the first option, frequencies were simply calculated from the base composition. The second option made use of the dinucleotide frequencies to take into account the biased dinucleotide composition of DNA sequences (Nussinov 1984). The frequencies of single-base runs of the 3′UTR were higher than the theoretical frequencies calculated by either method. However, it is apparent that doublet frequencies gave higher theoretical values than singlet frequencies, indicating that the bias in dinucleotide composition contributes to the abundance of single-base repeats at the 3′UTR. The effect is due to the overrepresentation of all four single-base dinucleotides, in agreement with the rule for most eukaryotic sequences (Nussinov 1981; Karlin and Burge 1995).

Single-base repeats are generally conserved between the mouse, the rat, and humans, but as shown in Fig. 1, they often differ in length by one or two nucleotides. Comparisons of the 3′UTRs from rat PC12 cells and brain shows the presence of nucleotide substitutions and insertions/deletions (Fig. 1). Insertions/deletions are of a single nucleotide and are always associated with single-base runs. This fact further suggests that single-base runs could be a favored substrate of slippage mechanisms.

Two-Base Runs

Other simple sequences present in the 3′UTR are based on two different nucleotides. The simplest forms of this kind of sequence are dinucleotide repeats such as $(TA)_n$. Most frequently, each nucleotide is reiterated a variable number of times, as in $T_2AT_5A_2TA_4TA$, so that no regularity is apparent. Examples of this kind of two-base runs belonging to the homopurine, homopyrimidine, and purine-pyrimidine classes are shown in Fig. 5.

Sequence Simplicity

We have estimated the clustering of simple motifs present either in short arrays or interspersed along the sequence with SIMPLE34 (Hancock and Armstrong 1994), a modified version of the SIMPLE program (Tautz et al. 1986). SIMPLE34 allows the comparison of the test sequences with random sequences generated from either the base composition or the dinucleotide composition of the test sequence. The latter option takes into account the biased dinucleotide composition of DNA sequences (Nussinov 1984).

Figure 6 shows the simplicity profiles for the 3′UTR from the rat and humans. Comparison of the SFs of the rat and human sequences with the SFs of the random sequences obtained from the base composition gave RSFs of 1.139 ($P < 0.05$) and 1.542 ($P < 0.003$), respectively. When doublet frequencies were used in the generation of the random sequences, the RSF of the rat sequence became nonsignificant (1.005), while the RSF of the human sequence decreased to 1.332, but still retaining the same confidence level ($P < 0.003$). A main contribution to the overall simplicity of the 3′UTR is made by single-base repeats and two-base runs. At the local level, a large proportion of peaks in the simplicity profiles is also caused by motifs of this kind.

The higher values of the SFs of the random sequences calculated using the doublet frequencies instead of the base composition indicate that the biased dinucleotide composition of the 3′UTR contributes to the clustering of sequence motifs. It has been pointed out that genes with a more biased base composition should also contain a higher frequency of clustered motifs (Hancock 1995a). In the 3′UTR noncoding strand, G and T are more abundant than A and C. Therefore, the biases in both nucleotide and dinucleotide composition presumably contribute to the clustering of motifs that could act as substrates for slippage.

pirymidine runs

$CTC_2T_2C_2TC_3TC_2T$
CTC_3TC_2T
$T_3CTCTCT_2$
$T_4CTC_2T_6CTCTCT_2$
CT_4CTCT_2CT
$T_3CT_5CTCT_2CT$
$T_3C_2T_2CTC$
$T_3C_2T_3C_4TCTC_2T$
$CT_4CT_4CT_4C_2TCT$
TCT_3CTCT_2

purine runs

G_4AGA_2GA
$AG_2AGAGAG_2$
$A_2G_2AG_3AG_3A_3$
$A_3G_2A_2G_2AG_2$
$G_4AGA_2GAGA_2$
$AG_3A_2G_4AG$
$GAG_2AG_4AGAG_2$

purine-pirymidine runs

$(TA)_3$
$(TA)_4$
$T_3AT_4A_2TA_3$
$T_2AT_5A_2TA_4TA$
$GTGTGT_3$
$G_2TG_2T_3GT_3$
$(GT)_3$
$(GT)_4$
$T_3G_2T_3G_2TG_4$
$T_2G_2TGTGT_7G_2$

Fig. 5. Examples of homopurine, homopyrimidine, and purine-pyrimidine sequences present in the 3′UTR of the H1° gene from the mouse, rat, and human sequences. Runs of a single base are not included.

Discussion

The rates of nucleotide substitution and divergence in 3′UTRs indicate that these sequences are under low selective pressure (Efstratiadis et al. 1980; Miyata et al. 1980; Li et al. 1985). Constraints on 3′UTRs may operate mainly at the level of the secondary and tertiary structures of the mRNA and also locally, on sequences involved in protein binding (Christ and Nath 1993). It has been proposed that regions that are evolving rapidly could be undergoing coevolution either with binding proteins or within the mRNA secondary structure (Hancock and Dover 1990). In genes with long 3′UTRs, length

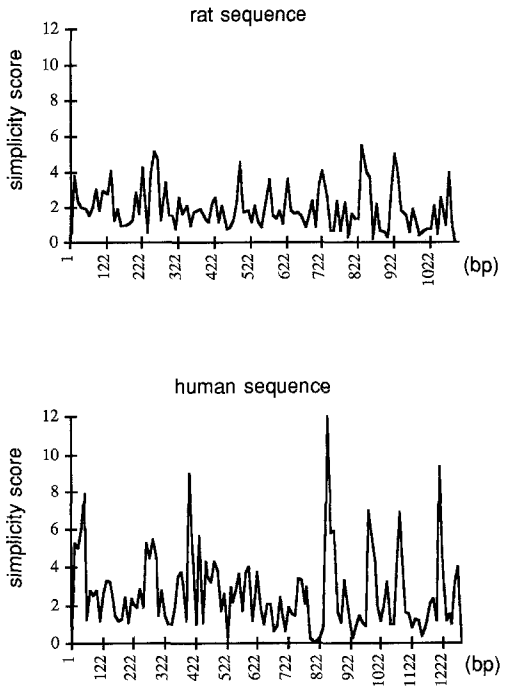

Fig. 6. DNA sequence simplicity profiles of the 3′ untranslated region of the H1° gene from the rat and humans. Profiles were generated with SIMPLE34 (Hancock and Armstrong 1994). The horizontal axis represents the length of the sequence and the vertical axis the simplicity score. The scores from the simplicity factor calculation were averaged in ten nucleotide steps.

expansion may have been positively selected for reasons not yet understood, but probably related to mRNA stability and posttranscriptional regulation (Rousseau et al. 1992; Brocard et al. 1994). The 3′UTRs of H1° of the rat and humans have a length difference of 185 nucleotides. This phenomenon recalls the length mutations of simple sequence loci associated with human genetic diseases (Caskey et al. 1992; Bates and Lehrach 1994; Richards and Sutherland 1994), although in the case of the 3′UTR the length difference has no apparent functional effect. Constraints are not uniformly distributed along the 3′UTR of H1°, as judged by the differences in local conservation. In short 3′UTR, the downstream portion is generally more conserved than the more upstream sequences, probably due to functional requirements (Miyata et al. 1980). In the 3′UTR of H1°, three main regions of homology can clearly be distinguished in dot matrix comparisons: a highly variable central region (69% homology), flanked by more conserved regions (78% and 80% homology). The central region is not only more variable, but it also absorbs about 60% of the length difference between the rat and human sequences. The regions of mRNA sequences involved in secondary structures are generally more conserved (Gutell 1993). The low degree of conservation of the central region the 3′UTR of H1° suggests the possibility that it could act as a hinge between the more conserved flanking regions.

Pairwise comparisons of the 3′UTRs of the mouse, the rat, and humans reveal that these regions have diverged by insertion/deletion and by nucleotide substitution. Insertions/deletions generally have one to 20 nucleotides. Direct repeats of 3–7 bp are often found near the ends of insertions/deletions. Direct repeats promote insertions and deletions by slippage mechanisms during DNA replication both *in vivo* (Levinson and Gutman 1978b; Henderson and Petes 1992; Strand et al. 1993) and *in vitro* (Schlötterer and Tautz 1992). Direct repeats were previously found flanking known deletions in the coding and noncoding regions of mammalian β-like globin genes (Efstratiadis et al. 1980). Some gaps in the 3′UTR of H1° do not present direct repeats in their vicinity. In sequences under low selective pressure, such as the 3′UTRs, direct repeats are not indeed expected in association with the totality of insertions/deletions, since nucleotide substitutions and the possible repeated occurrence of slippage events nearly at the same place should have eliminated the repeats in a number of cases.

One of the most apparent features of the 3′UTR of H1° is the abundance of single-base repeats. This phenomenon was previously reported in intron sequences, which are considered to be under low selective pressure, by Blaisdell (1983) and Levinson and Gutman (1987a), and taken as an argument in favor of the ubiquity of slipped-strand mispairing processes. We have extended this analysis to the 3′UTR of histone H1°. The frequencies of single-base repeats in the natural sequences are higher than the theoretical frequencies calculated from either the base or the dinucleotide compositions. The differences become larger with increasing size category. Pairwise comparisons have shown that although single-base repeats are generally conserved in the mouse, rat, and humans, they often differ in length by one or two nucleotides. A large proportion of the gaps of 1 or 2 nucleotides results, in fact, from length differences in single-base repeats. Such length differences in conserved single-base repeats are an additional argument in favor of the active involvement of slippage processes in the evolution of single-base repeats.

Other simple motifs characteristic of the 3′UTR of H1° are two-base runs. Homopurine, homopyrimidine, and purine-pyrimidine runs are present. In these sequences, either nucleotide is reiterated a variable number of times so that no regularity is apparent. Such sequences are potential substrates for slippage events, involving both contiguous and noncontiguous repeats. Two-base runs could evolve either toward an increase in overall simplicity through the insertion of new simple motifs and deletion of nonrepetitive sequences, or toward a regain of complexity by deletion of repeats and base substitutions. Two-base runs can be understood in terms of a dinucleotide motif, which has been made cryptic by the nonregular reiteration of either nucleotide. The mechanisms of replication slippage could have contributed to the early evolution of these sequences through the amplification of a tandemly repeated dinucleotide motif and then, in com-

bination with point mutations, expanded the sequence with single-base runs of either nucleotide.

Comparison of the 3′UTR from PC12 cells reported here and from rat brain has shown the presence of several nucleotide substitutions and insertions/deletions. Several mismatches were also reported between the 3′UTR from mouse teratocarcinoma and from murine erythroleukemia cells (Brocard et al. 1994). These sequence differences could reflect a polymorphism in the 3′UTR, as predicted by the neutral mutation hypothesis for regions that evolve at high rates (Hudson et al. 1987).

We have analyzed the simplicity of the 3′UTR with the program SIMPLE34 (Hancock and Armstrong 1994). The human sequence is significantly simpler than the rat sequence. Apparently, slippage products have been more actively incorporated to the human lineage than to the rodent lineage. This is in contrast with the fact that rodent microsatellites are generally longer than human microsatellites (Beckmann and Weber 1992; Love et al. 1990), but is consistent with the expansion of microsatellite loci in humans relative to their homologous counterparts in other primates (Rubinsztein et al. 1995). It also agrees with the accumulation of simple sequences as genome size increases (Hancock 1995b). The higher simplicity and length of the human sequence could then be at the same time a reflection of a general tendency and a peculiarity of the human lineage (Dover 1995).

Sequence simplicity and variation have been correlated in regions of the ribosomal genes that have evolved by insertional mutagenesis (Tautz et al. 1986) and in some regions of the locus *hunchback* of *Drosophila* (Treiter et al. 1989). In the 3′UTR, the simplicity peaks are distributed rather uniformly along the sequence, which is consistent with a relatively uniform distribution of insertions/deletions.

In summary, several features provide evidence to support the involvement of replication slippage in the evolution of the 3′UTR of the H1° gene. These features include (1) the length imbalance of the rat and human sequences; (2) the presence of numerous short (<20 bp) insertions/deletions, which are often flanked by direct repeats; (3) the abundance of single-base repeats, two-base runs, and other simple motifs; (4) the association of short (1–2 bp) gaps with length differences in conserved single-base repeats; and (5) the presence of single-base length polymorphisms in the rat and mouse sequences.

Acknowledgments. We thank Prof. A. Fontdevila and Dr. M. Santos for reading the manuscript and for suggestions. This work was supported by the CICYT (PB92-0633 and PB93-0369) and the Generalitat de Catalunya (CIRIT, Grups de Recerca de Qualitat, GRQ93-2063).

References

Albig W, Drabent B, Kunz J, Kalff-Suske M, Grzeschik K-H, Doenecke D (1993) All known human H1 histone genes except the H1° gene are clustered on chromosome 6. Genomics 16:649–654

Alonso A, Breuer B, Bouterfa H, Doenecke D (1988) Early increase of histone H1° mRNA during differentiation of F9 cells to parietal endoderm EMBO J 7:3003–3008

Bates G, Lehrach H (1994) Trinucleotide repeat expansions and human genetic diseases. Bioessays 4:277–284

Beckmann JS, Weber JL (1992) Survey of human and rat microsatellites. Genomics 12:627–631

Blaisdell BE (1983) A prevalent persistent global nonrandomness that distinguishes coding and non-coding eukaryotic nuclear DNA sequences. J Mol Evol 19:122–133

Brocard M-P, Rousseau D, Lawrence JJ, Khochbin S (1994) Two mRNA species encoding the differentiation-associated histone H1° are produced by alternative polyadenylation in mouse. Eur J Biochem 221;421–425

Caskey CT, Pizzuti A, Fu Y-H, Fenwick RG, Nelson DL (1992) Triplet repeat mutations in human disease. Science 256:784–789

Castiglia D, Gristina R, Scaturru M, Di Liegro I (1993) Cloning and analysis of a cDNA for rat histone H1°. Nucleic Acids Res 21:1674

Costa AR, Peixoto AA, Thackeray JR, Dalgleish R, Kyriacou CP (1991) Length polymorphism in the thereonine-glycine encoding repeat regions of the period gene in Drosophila. J Mol Evol 32: 238–246

Christ B, Nath A (1993) The glucagon-insulin antagonism in the regulation of cytosolic protein binding to the 3′ end of phosphoenolpyruvate carboxykinase mRNA in cultured rat hepatocytes. Eur J Biochem 215:541–547

Decker CJ, Parker P (1995) Diversity of cytoplasmic functions for the 3′ untranslated region of eukaryotic transcripts. Curr Opin Cell Biol 7:386–392

Djian P, Green H (1989) Vectorial expansion of the involucrin gene and the relatedness of hominoids. Proc Natl Acad Sci USA 86: 8447-8451

Doenecke D, Tönjes R (1986) Differential distribution of lysine and arginine residues in the closely related histones H1° and H5: analysis of a human H1° gene. J Mol Biol 187:461–464

Dover G (1995) Slippery DNA runs on and on. . . . Nat Genet 10:254–256

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, Deriel JK, Forget BR, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot J (1980) The structure and evolution of the human β-globin gene. Cell 21:653–668

Eickbush TH, Burke WD (1986) The silkmoth late chorion locus. J Mol Biol 190:357–366

García-Segura LM, Luquín S, Martínez P, Casas MT, Suau P (1993) Differential expression and gonadal hormone regulation of histone H1° in the developing and adult rat brain. Dev Brain Res 73:63–70

Gjerset R, Gorka C, Hasthorpe S, Lawrence JJ, Eisen H (1982) Developmental and hormonal regulation of protein H1° in rodents. Proc Natl Acad Sci USA 79:2333–2337

Gutell RR (1993) Comparative studies on RNA: inferring high-order structure from patterns of sequence variations. Curr Opin Structural Biol 3:313–322

Hancock JM (1995a) The contribution of DNA slippage to eukaryotic nuclear 18S rRNA evolution. J Mol Evol 40:629–639

Hancock JM (1995b) The contribution of slippage-like processes to genome evolution. J Mol Evol 49:1038–1047

Hancock JM, Armstrong JS (1994) SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Comput Appl Biosci 10:67–70

Hancock JM, Dover GA (1988) Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. Mol Biol Evol 5:377–391

Hancock JM, Dover GA (1990) Compensatory slippage in the evolution of ribosomal RNA genes. Nucleic Acids Res 18:5949–5954

Heintz N, Zernik M, Roeder RG (1981) The structure of the human histone genes: clustered but not tandemly repeated. Cell 24:661–668

Henderson ST, Petes TD (1992) Instability of simple sequence DNA in *Saccharomyces cerevisiae*. Mol Cell Biol 12:2749–2757

Hentschell CC, Birnstiel ML (1981) The organization and expression of histone gene families. Cell 25:301–313

Hoelzel AR, Hancock JM, Dover GA (1991) Evolution of the Cetacean mitochondrial D-loop region. Mol Biol Evol 8:475–493

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159

Izaurralde E, Käs E. Laemmli UK (1989) High preferential nucleation of histone H1 assembly on scaffold-associated regions. J Mol Biol 210:573–585

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283–290

Lafarga M, García-Segura LM, Rodríguez JR, Suau P (1995) Expression of histone H1° in transcriptionally activated supraoptic neurons. Mol Brain Res 29:317–324

Lennox RW, Cohen LH (1983) The histone H1 complements of dividing and nodividing cells of the mouse and their interphase phosphorylated states in differentiated and undifferentiated cell lines derived from murine teratocarcinomas. J Biol Chem 258:262–268

Levinson G, Gutman GA (1987a) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Levinson G, Gutman GA (1987b) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Res 15:5323–5338

Li W-H, Luo C-C, Wu C-I (1985) Evolution of DNA sequences molecular. In: Ross J, Macintery J, (eds) Evolutionary genetics. Plenum Press, New York, pp 1–84

Love JM, Knight AM, McAleer MA, Todd JA (1990) Towards construction of a high resolution map of the mouse genome using PCR analyzed microsatellites. Nucleic Acids Res 18:4123–4130

Martínez P, Vidal JM, Monsalves C, Pucket C, Ponte I, Suau P (1995) Cloning and analysis of the coding region of the histone H1°-encoding gene from rat PC12 cells. Gene 166:313–316

Meersseman G, Pennings S, Bradbury EM (1991) Chromatosome positioning on assembled long chromatin. Linker histones affect nucleosome placement on 5S rDNA. J Mol Biol 220:89–100

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Nussinov R (1981) Nearest neighbor nucleotide patterns. Structural and biological implications. J Biol Chem 256:8458–8462

Nussinov R (1984) Strong doublet preferences in nucleotide sequences and DNA geometry. J Mol Evol 20:111–119

Palazzolo MJ, Meyerowtiz EM (1987) A family of lambda phage cDNA cloning vectors, λ5WAJ, allowing the amplification of RNA sequences. Gene 52:197–206

Paulsson G, Lendahl U, Galli J, Ericsson C, Wieslander L (1990) The balbiani ring 3 gene in *Chiromonus tetans* has a diverged repetitive structure split by many introns. J Mol Biol 211:331–349

Ponte I, Martínez P, Ramírez A, Jorcano J, Monzó M, Suau P (1994) Transcriptional activation of histone H1° during neuronal terminal differentiation. Dev Brain Res 80:35–44

Richards RI, Sutherland GR (1994) Simple repeat DNA is not replicated simply. Nat Genet 6:114–115

Rousseau D, Khochbin S, Gorka C, Lawrence JJ (1992) Induction of H1° gene expression in B16 murine melanoma cell. Eur J Biochem 208:775–779

Rubinsztein DC, Amos W, Leggo J, Goodburn S, Jain S, Li S-H, Margolis RL, Ross CA, Ferguson-Smith MA (1995) Microsatellite evolution—evidence for directionality and variation in rate between species. Nat Genet 10:337–343

Schlötterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. Nucleic Acids Res 20:211–215

Seyedin SM, Kistler WS (1979) H1 sub-fraction of mammalian testes. I. Organ specificity of the rat. Biochemistry 18:1371–1375

Smith TF, Waterman MS (1981) Comparison of bio-sequences. Adv Appl Math 2:482–489

Strand M, Prolla TA, Liskay RM, Petes TD (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365:274–276

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Tautz D, Schlötterer C (1994) Simple sequences. Curr Opin Genet Dev 4:832–837

Trieter M, Pfeifle C, Tautz D (1989) Comparison of the gap segmentation gene hunchback between *Drosophila melanogaster* and *Drosophila viridis* reveals novel models of evolutionary change. EMBO J 8:1517–1525

Zlatanova, Donecke (1994) Histone H1°: a major player in cell differentiation. FASEB J 8:1260–1268