'Compensatory slippage' in the evolution of ribosomal RNA genes

John M.Hancock and Gabriel A.Dover
Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

Received August 13, 1990; Revised and Accepted September 28, 1990

ABSTRACT

The distribution patterns of shared short repetitive motifs in the expansion segments of the large subunit rRNA genes of different species show that these segments are coevolving as a set and that in two examined vertebrate species the RNA secondary structures are conserved as a consequence of runs of motifs in one region being compensated by complementary motifs in another. These unusual processes, involving replication-slippage, have implications for the evolution of ribosomal RNA and for the use of the rDNA multigene family as a 'molecular clock' for assessing relationships between species.

INTRODUCTION

Eukaryotic ribosomal RNA genes, in particular the large subunit rRNA (LSU-rRNA) genes, are mosaic structures of conserved 'core' segments and hypervariable 'expansion' segments. The former are defined by their homologous counterparts in prokaryote ribosomal RNA genes and are considered to be largely essential for ribosome functions [1,2]. They evolve slowly and evenly, and have provided a useful molecular metronome for assessing relationships between organisms stretching back to the inception of life's major Kingdoms [3,4]. In contrast, the expansion segments (sometimes also called variable regions) have no precise equivalents in prokaryotes [2,5] and reveal high variability in primary sequence within and between species [6–9], seemingly as a consequence of DNA slippage-like processes [9–11]. RNA secondary structure modelling shows that expansion segments from highly divergent taxa can adopt secondary structures which are similar despite having substantially divergent primary structures [10–13]. As is the case with the core segments, this could indicate a conserved function.

The origins and functions of expansion segments are intriguing in the light of recent findings which show that in some organisms there is no essential requirement for any particular linear arrangement of 'core' and 'expansion' segments within the LSU- and SSU- (small subunit) rRNA genes for either transcription or correct patterns of RNA secondary folding [14–16]. It is still uncertain whether expansion segments and the various spacers separating the genes (in particular the two internal transcribed spacers ITS1 and ITS2), are remnants of much longer ancestral sequences that once separated ribosomal minigenes, or whether they represent recent insertions of extraneous sequences (for recent discussion see refs 14–17). This uncertainty parallels the 'early-or-late' debate of intron evolution within RNA polymerase II transcribed, protein-coding genes [18–20].

In order to assess the mutational forces that have played on the expansion segments and to clarify their structural consequences at the DNA and RNA levels, we have made a detailed analysis of these segments from two vertebrate species. This analysis combines various algorithms for the detection of both fine-grained non-random patterns of sequence [21] and the distribution of short repetitive motifs (see below) with models of ribosomal RNA folding. With these procedures, we have identified a novel phenomenon which we call 'compensatory slippage'.

MATERIALS AND METHODS

Sequence analysis. Dot matrix analysis of sequences of *Homo sapiens* 28S rRNA [22] and *Xenopus laevis* 28S rRNA [23] was carried out using the DIAGON program [24]. The analysis of internal repetition of short sequence motifs within sequences (simplicity analysis), including generation of relative simplicity values and simplicity profiles, was carried out using the SIMPLE program [21] with minor modifications [JMH, unpublished]. Displays of distributions of trinucleotide motifs within sequences were produced using DIAGON [24] as follows: the sequence in question was compared to a file containing blocks of fifteen of each of the 64 possible trinucleotide motifs (from AAA to TTT). Trinucleotides were interspersed with hyphens to eliminate spurious positives (i.e. the file was of the form AAA-AAA-AAA- and so on). Comparisons were carried out at a stringency of 3 out of 3. Chi-squared analysis of trinucleotide composition was carried out using a TurboBasic program (THREES) run under DOS 3.30 on an IBM PS/2 microcomputer. This program calculated chi-squared values using expected frequency values calculated from the base composition of the sequence. Minimum energy RNA secondary structures were calculated using the FOLD algorithm [25] implemented under the UWGCG package [26].

RESULTS AND DISCUSSION

Coevolution of Simple Sequence Expansion Segments

Programs [21] that look for internal repetition and sequence simplicity show that expansion segments are internally repetitious [9], with high levels of a property known as cryptic simplicity [21], that is, scrambled permutations of short repetitive motifs within defined regions. DNA simplicity probably reflects the

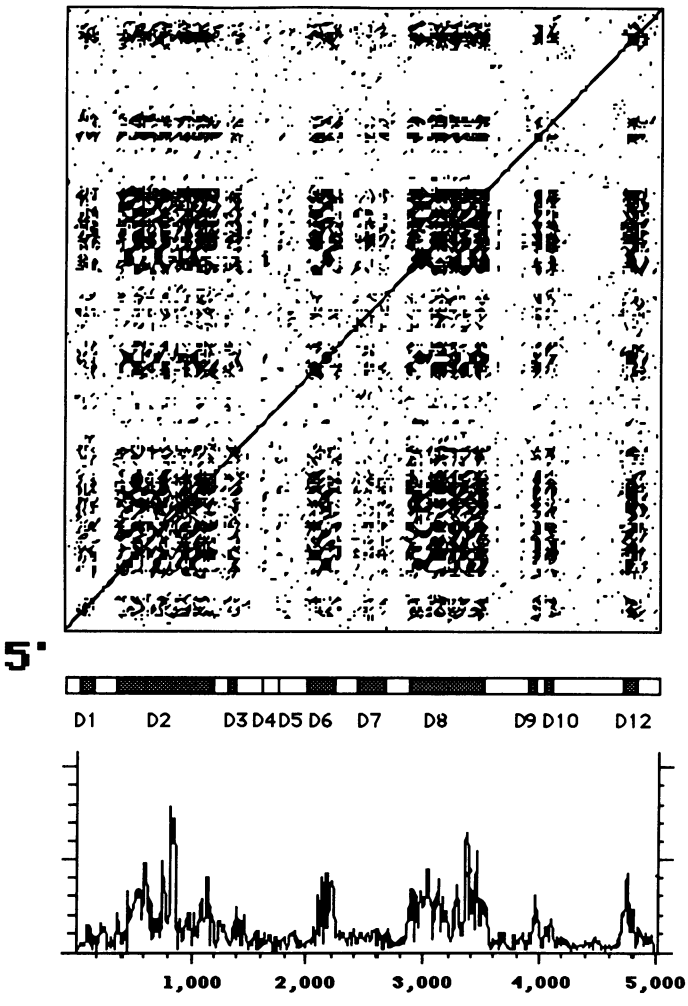

Figure 1. Dot matrix and sequence simplicity analysis of the human 28S rRNA gene. The dot matrix was generated using the DIAGON program [24]. Both axes correspond to the length of the gene, with the 5' end at the bottom left hand corner. The matrix corresponds to a comparison of all points within the sequence to all other points. To visualize regions of repetition corresponding to expansion segments, the analysis was carried out so that any two stretches of sequence 35 nucleotides long which matched at any 19 or more of those 35 nucleotides is represented as a single dot. The positions of the expansion and core segments within the sequence are marked below the dot matrix, with expansion segments being represented as shaded boxes and labelled D1–D12. At the bottom of the figure is a display of the level of sequence simplicity (as measured by the SIMPLE program [21,9], see text) at all points along the sequence (simplicity profile). Expansion segments can be seen to correspond to regions of internal repetition within the dot matrix, and to regions of high sequence simplicity.

products of a slippage-like mechanism of DNA turnover [21]. A visual display (simplicity profile) of sequence simplicity in a human 28S rRNA gene is shown in Fig.1 (for details, see legend). The dot matrix analysis (Fig.1) also reveals that the expansion segments are similar to each other. Interestingly, species like yeast, nematode and slime-mold, whose expansion segments are not cryptically simple, do not have expansion segments which are similar in sequence. Hence co-evolution of expansion segments is intimately involved with slippage-like mutational mechanisms.

Distribution Patterns of Simple Sequences

In order to understand the effects of slippage on RNA folding patterns within the expansion segments, we have analysed the precise patterns of distribution of simple sequence motifs in LSU-rRNA expansion segments of vertebrate species which, along with those of rice (*Oryza sativa*), show the strongest patterns of expansion segment co-evolution [9]. To simplify the analysis we have concentrated on two expansion segments, D2 and D8 (using the numbering system of Hassouna *et al* [27]), which are two of the most prominent both with respect to the densities of similarities in dot matrix analyses and in the degrees of sequence simplicity [9,13] (see Fig.1 for example). Analysis was carried out at the level of trinucleotide repeats, allowing visualization of the precise sequence motifs responsible for the high levels of sequence simplicity in the expansion segments and their fine-grained distribution (see Materials and Methods and Figure 2a and legend for the operation of this procedure).

Figure 2b shows the distribution of all possible trinucleotide motifs within the human 28S rRNA gene. Regions of the sequence corresponding to the conserved core regions and expansion segments are indicated below the diagram by open and shaded bars, respectively. The plot shows that the trinucleotide composition of human 28S rRNA gene core sequences is close to random but that the expansion segments show highly non-random distributions of trinucleotide motifs. Figure 2c, which is an analysis of nucleotides 2874–4020 of the same gene, shows that the boundaries of the expansion segments (in this case expansion segments D8 and D9) are characterized by sharp changes in the trinucleotide composition of the sequence. Figures 2c and 2d (which is a higher resolution analysis of expansion segment D2) show that the trinucleotide composition also varies within the expansion segment sequences themselves. Particular parts of the sequence of each expansion segment are characterized by the preferential presence of only a small subset of those trinucleotides which are overrepresented within the expansion segment as a whole. For example, one region of expansion segment D8 (Figure 2c) contains long runs of three or four trinucleotide motifs: CGG, GCG, GGC, and to a lesser extent GGG. Another region within D8 is composed almost entirely of tandemly repeated CCC triplets. Similar regions of high concentrations of specific trinucleotide motifs can be found in expansion segment D2 (Fig. 2d) and in other highly simple expansion segments. Detailed inspection of the distribution patterns of specific motifs show that they can often occur in mutually exclusive blocks. These are shown as shaded regions within D8 and D2 (Fig.2c & d).

To assess the statistical significance of the trinucleotide composition of expansion segments, we carried out chi-squared analysis of the trinucleotide frequencies in the expansion segments of *Homo sapiens* [23] and *Xenopus laevis* [24] (Table 1). These two species both have GC-rich expansion segments which show similarity across species in dot-matrix analysis [9]. The table lists chi-squared values for the frequencies of all trinucleotide motifs within each expansion segment compared to expected values calculated on the basis of base composition. This analysis shows that those expansion segments which have strong patterns of self- and cross-similarity within a species, and which appear as prominent peaks in simplicity profiles (Fig.1) [9], also have highly significant chi-squared values ($p < 0.001$), indicating that they have highly non-random trinucleotide compositions. The table also lists the most highly represented trinucleotide motifs within individual expansion segments. It is clear that the most cross-related expansion segments share many highly repeated trinucleotide motifs. In human 28S rDNA, for example, expansion segments D1, D2, D4, D6, D8, D9 and D12 share

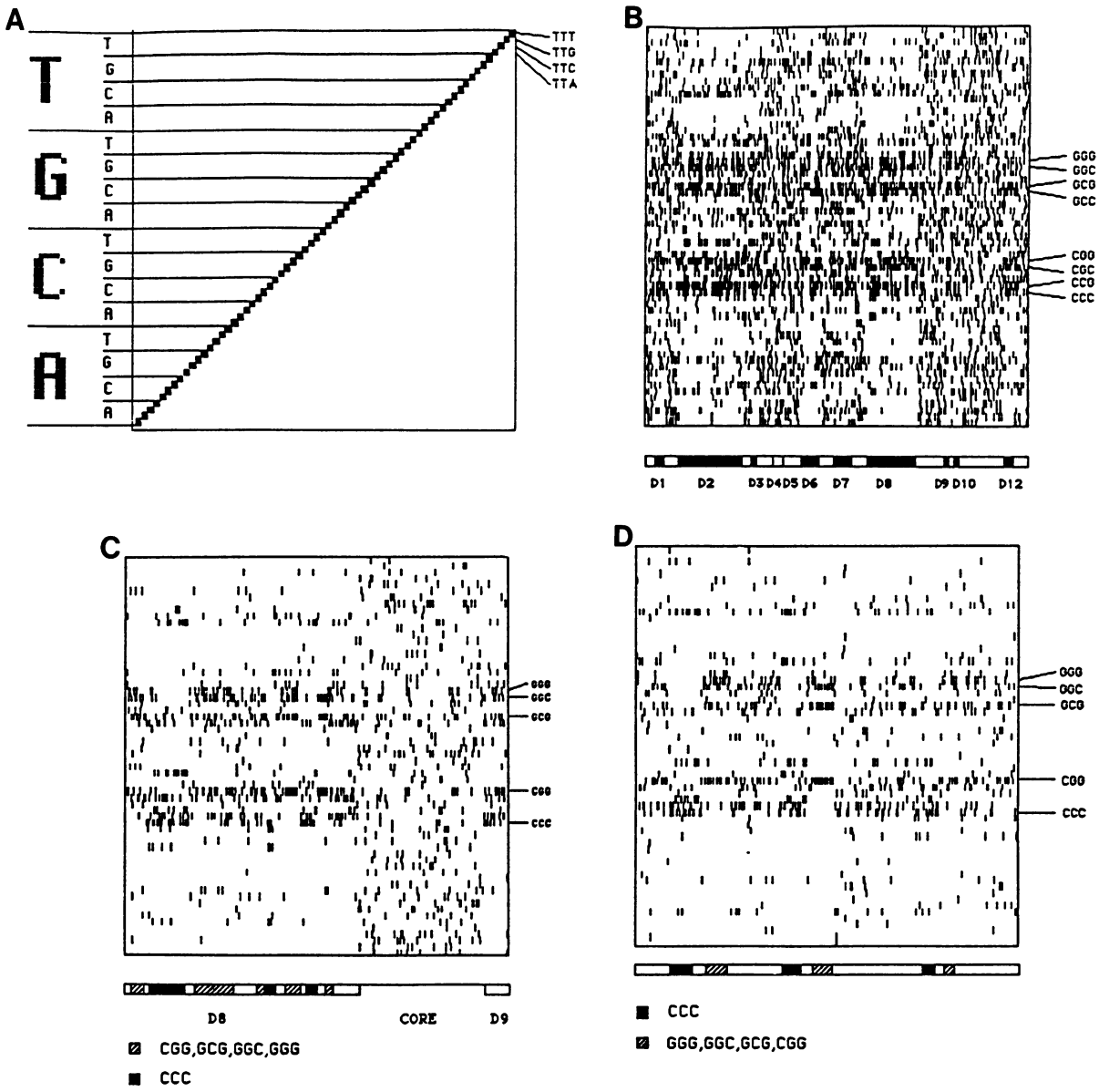

Figure 2. Distributions of trinucleotide motifs within human 28S rRNA sequences. Sequences are compared to a file containing tandem arrays of all 64 possible trinucleotide motifs using the DIAGON program [24] at a stringency of 3 out of 3. Trinucleotide motifs (corresponding to horizontal rows of vertical bars) are arranged in alphabetical order from AAA (bottom) to TTT (top). A: Arrangement of trinucleotide motifs within the frame (vertical axis). The largest subdivisions (large letters) represent the first letter of the trinucleotide motif, the second subdivision represents the second letter. The second subdivision is divided again by four, representing A, C, G or T at the third triplet position. The positions of triplets TTA, TTC, TTG and TTT are labelled individually as an illustration. The diagonal shows the location of each motif within the artificial probe sequence. B: Distribution of trinucleotide motifs within the complete human 28S rRNA gene. Trinucleotide motifs which are strongly overrepresented in expansion segments are labelled. The positions of expansion segments within the sequence are indicated by shaded boxes, and expansion segments labelled individually (D1 – D12). C: High resolution display of the distribution of trinucleotide motifs for nucleotides 2874 – 4020 of the human 28S rRNA gene. Highly represented trinucleotides and expansion segments are identified as in A. D: High resolution display of trinucleotide distribution within human expansion segment D2. Highly represented trinucleotides are labelled. Shaded and hatched bars at the bottom of 2 c & d correspond to regions which contain high concentrations of CCC (shaded) or of GGG, GGC, GCG and CGG (hatched).

high representation of the motif CCC, whilst D6, D7b, D8 and D12 share high levels of GCG/CGC. It is also noticeable that certain motifs, and in particular CCC, are highly represented in the same expansion segment in both species. For example, CCC is overrepresented in D1, D2, D8 and D9 of both the *X. laevis* and the human LSU-rRNA gene.

This analysis confirms the results of motif distribution analysis (Fig.2) that certain trinucleotide repeats are highly concentrated in expansion segments which are highly repetitive and which show sequence similarity to one another. This provides a basis for our observation of co-evolution between expansion segments, in that these expansion segments all appear to undergo a slippage-like process in such a way as to accumulate similar trinucleotide motifs.

RNA Secondary Structures and 'Compensatory Slippage'

Fig. 3 (a – c) shows predictions of the secondary structures of three expansion segments: human D2 and D8 and *X. laevis* D2, generated using the algorithm of Zuker & Stiegler [25,26]. Superimposed upon the secondary structures for human D2 and

Table 1. Triplets analysis of human and *X. laevis* expansion segments. Against each expansion segment are listed the following: 1) a chi-squared value (60 degrees of freedom) representing the extent to which the trinucleotide composition of the sequence deviates from that of a random sequence of the same length and base composition; 2) the level at which the calculated chi-squared value is significant (NS = not significant; 0.01 = significant at the $p<0.01$ level, and so on); and 3) the trinucleotide motifs which are most strongly represented within the sequence. D1−D12 are the numbered expansion segments starting from the 5′ end of the 28S rRNA gene. No analysis is presented for expansion segment D11, which is commonly less than 10 nucleotides long.

	HUMAN			XENOPUS		
ES	χ^2	p	MOTIFS	χ^2	p	MOTIFS
D1	49.41	NS	ccc	96.57	0.01	ccc,cgg
D2	**468.45**	0.001	ccc,ggg	**350.80**	0.001	ccc
D3	72.66	NS	ccg	**107.50**	0.001	ccc,ggg
D4	76.18	NS	ccc	64.12	NS	gtc
D5	46.03	NS	tgg	58.02	NS	gct
D6	**141.75**	0.001	ccc,gcg	54.08	NS	
D7a	63.96	NS	cga	59.38	NS	cga,tcg
D7b	45.83	NS	cga,gcg	97.44	0.01	gcg,ggg
D8	**522.13**	0.001	ccc,gcg	**307.61**	0.001	ccc
D9	96.37	0.01	ccc	58.03	NS	ccc,cgc
D10	61.61	NS	ggg	74.89	NS	ccc
D12	**103.67**	0.001	ccc,cgc	94.71	0.01	

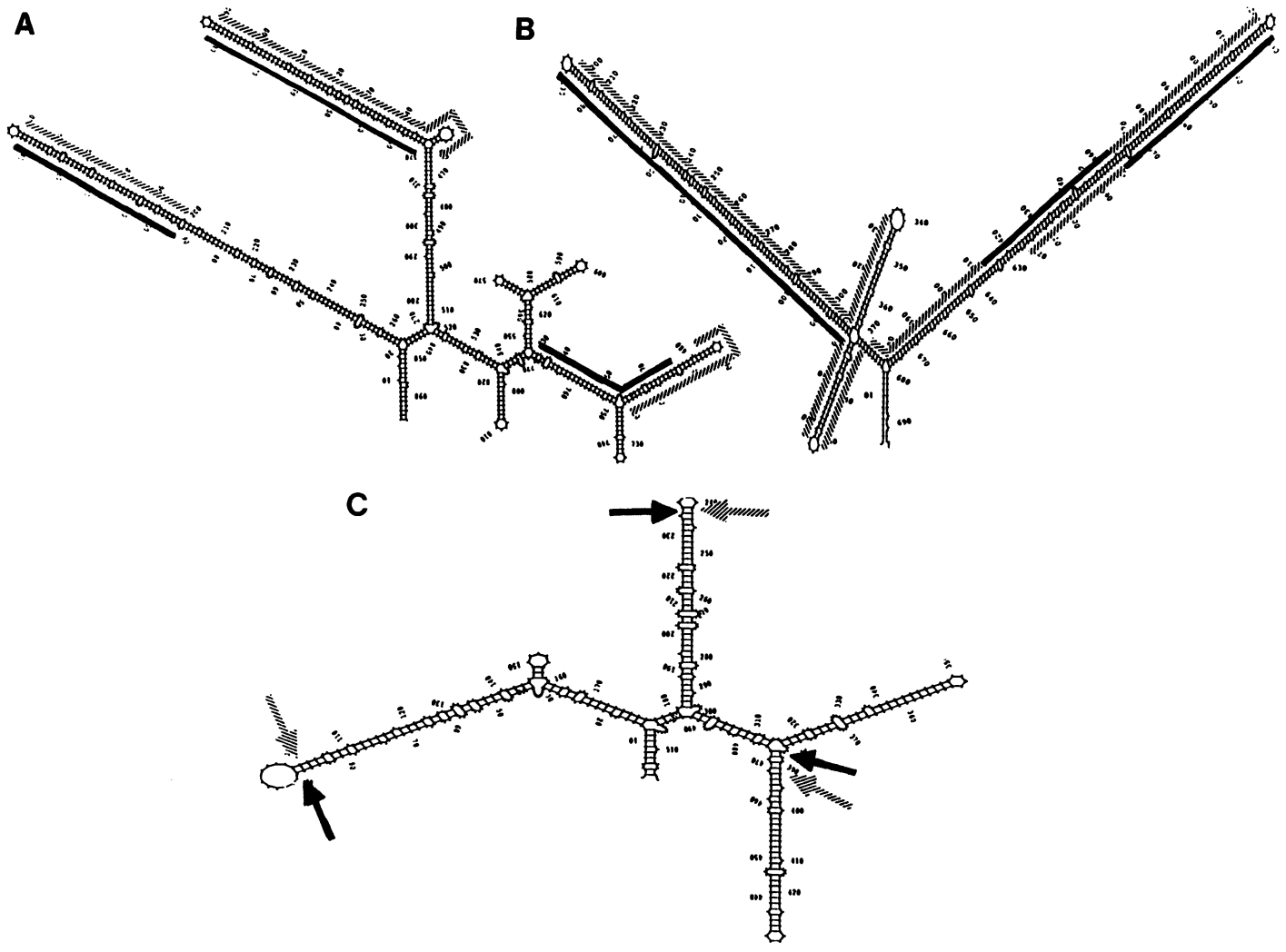

Figure 3. Secondary structure models of expansion segments. A: Human expansion segment D2; B: Human expansion segment D8; C: *X. laevis* expansion segment D2. Secondary structures were generated using the FOLD program [25] implemented under the GCG package [26] on the Cambridge University Molecular Biology VAX. Regions corresponding to simple sequence blocks identified in Figs 2 b−d are indicated by shading (CCC-rich) or hatching (GGG-rich). Arrows in C represent positions of insertion of simple sequence blocks in human D2 (see text). Their shading indicates the composition of the inserted sequences.

D8 are the locations of the blocks identified above as being rich in either CCC (shaded) or GGG/GCG (hatched). It is apparent that the mutually exclusive arrangement of these blocks along the primary sequence translates into an arrangement in the secondary structure whereby such blocks form complementary parts of stem-loop structures. Expansion segment secondary structures have seemingly resulted from the accumulation of slippage-generated products at sites within the DNA sequence such that a folding pattern ensues which corresponds to the overall architecture conserved between lineages. We term this process 'compensatory slippage'. The action of compensatory slippage is illustrated in Fig 3c, which shows the secondary structure of *X. laevis* expansion segment D2. The arrows, shaded or hatched to represent sequence composition in the same way as Figures 3a and b, indicate the likely positions within this expansion segment at which the blocks of C- and G-rich sequence, identifiable in the human sequence, appear to have inserted.

Compensatory Mutation and Compensatory Slippage

Compensatory slippage is consistent with previous observations that the expansion segments of highly divergent organisms adopt similar secondary structures despite having highly divergent sequences, [9,10,13] and that the major differences between long and short expansion segments lie in the lengths of certain major secondary structural stems [13]. As compensatory slippage reflects the insertion of products of slippage in such a way that they lie opposite one another in the rRNA secondary structure, it would seem likely that the localization of slippage-generated motifs along the rRNA gene sequence is restricted by natural selection to sites which allow them to pair stably without disrupting pre-existing rRNA secondary structures. The generation of such compensatory patterns of slippage-generated products presents an evolutionary paradox since some mechanism of 'crosstalk' must occur between regions which may be tens of nucleotides apart in the primary sequence. In order that compact secondary structures are maintained over long periods of evolutionary time such a mechanism must ensure that, during the high rate of production of such slippage- generated tracts, they are both similar in length and complementary in sequence.

The process of compensatory slippage is analogous to the process of compensatory point mutation, which occurs frequently within the conserved core regions of rRNA genes and is responsible for the preservation of conserved secondary structural elements despite the accumulation of point mutations [1−3]. The evolutionary dynamics of compensatory point mutation are complex: not only must both of the compensating mutations occur in one rDNA repeat unit, but these mutations must spread through the entire gene family, replacing previous copies of the gene [10].

The Evolution of Compensation in a Multigene Family

Given the multiple copy nature of the rDNA genes, the occurrence of a single point mutation within a single gene will probably have little effect on the viability of the individual in which it arises; it is essentially a neutral mutation when rare. Over time some such point mutations can be spread (molecular drive) concomitantly through the gene family and through the population (as a consequence of the homogenizing effects of unequal crossingover (and other turnover mechanisms) in the gene family [10,28,29]) up to a point at which the new, mutant gene might begin to have a negative effect on the viability of the population. Such inviability can be overcome with the occurrence and selection of a compensatory mutation that restores the correct rRNA folding.

A similar argument can be made to explain compensatory slippage. When slippage takes place at a particular site within a LSU-rRNA gene, a few nucleotides, say GGG, will be added to the sequence. This may result in a bulge in a stem. If such motifs continue to accumulate within the stem and are also being homogenized throughout the gene family, then a stage may be reached at which one or more stems are disrupted and natural selection may come into play to select against those individuals with the greatest numbers of such mutant repeats. Alternatively, selection may act on individuals which have accumulated copies of a second, complementary repeat (such as CCC) in such a position that it can pair with the first motif and preserve secondary structure. The probability that the new double mutant will spread will depend both on selection and on the rates and biases of stochastic processes in the gene family [28, 30].

Supporting evidence both for the action of slippage within expansion segments and for constraint on the number of slippage-generated products at any one position within a secondary structural stem comes from observations on sequence variability between expansion segments in human rRNA and rRNA genes [6−8] and from a detailed analysis of the progress of compensatory slippage and the molecular coevolution between expansion segments D2 and D8 in insect species (A. Ruiz Linares, JMH and GAD, in preparation).

Tempo of rDNA Evolution and the Molecular Clock

It is interesting that those species that show the most prominent accumulation of slippage-generated products in their expansion segments also have expansion segments with the most highly biased base compositions, namely the vertebrates and the monocotyledenous plant *O. sativa* (rice). Lesser degrees of detectable expansion segment similarity are present in Drosophila and in lemon (*Citrus limon*), which have less biased base compositions [9 and JMH, unpublished observations]. This may reflect the greater ease with which slippage-like mechanisms may be initiated in DNA sequences with biased base compositions, and therefore a higher concentration of repetitive motifs [31]. Hence, compensatory slippage will make a prominent contribution to the tempo of evolution of secondary structures in which it takes place.

Our observations that expansion segments are simple in sequence and coevolving, and that the incorporation of the products of slippage into rRNA genes is possibly constrained by the necessity to maintain rRNA secondary structure, might explain previously described anomalies of the rate of rDNA sequence divergence [32] and provide a cautionary tale in the use of rDNA as a 'molecular clock' for assessing species relationships. For example, the reliability of one particular method of phylogenetic reconstruction [33] has been questioned on the grounds that its analysis of LSU- and SSU-rRNA sequences of human, Drosophila, rice and Physarum places human and rice as a sister group, rather than the expected human-Drosophila grouping preferred by other methods [34]. However, our dot matrix [9] and trinucleotide motif comparisons show that human and rice LSU-rRNA sequences have, accidentally, converged on similar GC-rich motifs and, consequently, are more similar to each other than either is to Drosophila.

An Unanswered Question

Our analysis of the detailed motif composition of cryptic simplicity within LSU-rRNA genes has revealed the occurrence

of the process of compensatory slippage. This process may have played a role in the evolution of secondary structure in systems other than the rDNA. A major unanswered question in the case of rDNA is the means by which the coevolution of expansion segments of a given species has taken place. Why has the process of compensatory slippage made use of similar sequence motifs in different expansion segments? We are unable to distinguish between genomic explanations, involving a preference of slippage to use particular motifs as substrates no matter where they occur, and/or the occurrence of micro-gene-conversions between expansion segments, and explanations which involve selection.

REFERENCES

1. Noller, H.F. (1984) *Ann Rev Biochem* **53**, 119–162
2. Gerbi, S.A. (1985) in McIntyre, R.J. (ed) Molecular Evolutionary Genetics (Plenum, New York) pp 419–517
3. Woese, C.R. (1987) *Microbiol Rev* **51**, 221–271
4. Field, K.G., Olsen, G.J., Lane, D.J., Giavannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R. & Raff, R.A. (1988) *Science* **239**, 748–753
5. Clark, C.G. (1987) *J Mol Evol* **25**, 343–350
6. Gonzalez, I.L., Gorski, J.L., Campen, T.J., Dorney, D.J., Erickson, J.M., Sylvester, J.E. & Schmickel, R.D. (1985) *Proc Natl Acad Sci USA* **82**, 7666–7670
7. Maden, B.E.H., Dent, C.L., Farrell, T.E., Garde, J, McCallum, F.S. & Wakeman, J.A. (1987) *Biochem J* **246**, 519–527
8. Gonzalez, I.L., Sylvester, J.E. & Schmickel, R.D. (1988) *Nuc Acids Res* **16**, 10213–10224
9. Hancock, J.M. & Dover, G.A. (1988) *Mol Biol Evol* **5**, 377–391
10. Hancock, J.M., Tautz, D. & Dover, G.A. (1988) *Mol Biol Evol* **5**, 393–414
11. Hancock, J.M. & Dover, G.A. (1989) in A Fontdevila (ed) Evolutionary Biology of Transient Unstable Populations (Springer, Berlin) pp 206–220
12. de Lanversin, G. & Jacq, B. (1989) *J Mol Evol* **28**, 403–417
13. Michot, B. & Bachellerie, J.-P. (1987) *Biochimie* **69**, 11–23
14. Boer, P.H. & Gray, M.W. (1988) *Cell* **55**, 399–411
15. Boer, P.H. & Gray, M.W. (1988) *EMBO J* **7**, 3501–3508
16. Dover, G.A. (1988) *Nature* **336**, 623–624
17. Burgin, A.B., Parodos, K., Laue, D.J. & Pace, N.R. (1990) *Cell* **60**, 405–414
18. Gilbert, W. (1978) *Nature* **271**, 501
19. Doolittle, W.F. (1978) *Nature* **272**, 581–582
20. Rogers, J. (1989) *Trends In Genetics* **5**, 213–216
21. Tautz, D., Trick, M. & Dover, G.A. (1986) *Nature* **322**, 652–656
22. Gorski, J.L., Gonzalez, I.L. & Schmickel, R.D. (1987) *J Mol Evol* **24**, 236–251
23. Ware, V.C., Tague, B.W., Clark, C.G., Gourse, R.L., Brand, R.C. & Gerbi, S.A. (1983) *Nuc Acids Res* **11**, 7795–7817
24. Staden, R. (1982) *Nuc Acids Res* **10**, 4731–4751
25. Zuker, M. & Stiegler, P. (1981) *Nuc Acids Res* **9**, 133–148
26. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nuc Acids Res* **12**, 387–395
27. Hassouna, N., Michot, B. & Bachellerie, J.-P. (1984) *Nuc Acids Res* **12**, 3563–3583
28. Dover, G.A. (1982) *Nature* **299**, 111–117
29. Mian, A. & Dover, G.A. (1990) *Nuc Acids Res* **18**, 3795–3801
30. Dover, G.A. (1989) *Genetics* **122**, 249–252
31. Dover, G.A. & Tautz, D. (1986) *Phil Trans R Soc London B* **312**, 275–289
32. Woese, C.R. (1987) in Patterson, C. (ed) Molecules and Morphology in Evolution: Conflict or Compromise (Cambridge University Press, Cambridge, UK) pp 177–202.
33. Lake, J.A. (1987) *Mol Biol Evol* **4**, 167–191
34. Gouy, M. & Li, W.H. (1989) *Nature* **339**, 145–147