Gene, 54 (1987) 1–22
Elsevier

GEN 01990

The AT spacers and the *var1* genes from the mitochondrial genomes of *Saccharomyces cerevisiae* and *Torulopsis glabrata*: evolutionary origin and mechanism of formation *

(DNA sequence pattern; isostich analysis; polymorphism; divergence; intergenic sequences; *ori* sequences; untranslated sequences; intergenic open reading frames)

Miklos de Zamaroczy and Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 75005 Paris (France)

Received 25 July 1986
Revised 23 February 1987
Accepted 25 February 1987

SUMMARY

Intergenic sequences represent 63% of the mitochondrial 'long' (85 kb) genome of *Saccharomyces cerevisiae*. They comprise 170–200 AT spacers that correspond to 47% of the genome and are separated from each other by GC clusters, ORFs, *ori* sequences, as well as by protein-coding genes. Intergenic AT spacers have an average size of 190 bp, and a GC level of 5%; they are formed by short (20–30 nt on the average) A/T stretches separated by C/G mono- to trinucleotides. An analysis of the primary structures of all intergenic AT spacers already sequenced (32 kb; 80% of the total) has shown that they are characterized by an extremely high level of short sequence repetitiveness and by a characteristic sequence pattern; the frequencies of A/T isostichs conspicuously deviate from statistical expectations, and exponentially decrease when their (AT + TA)/(AA + TT) ratio, *R*, decreases. A situation basically identical was found in the AT spacers of the mitochondrial genome (19 kb) of *Torulopsis glabrata*.

The sequence features of the AT spacers indicate that they were built in evolution by an expansion process mainly involving rounds of duplication, inversion and translocation events which affected an initial oligodeoxynucleotide (endowed with a particular *R* ratio) and the sequences derived from it. In turn, the initial oligodeoxynucleotide appears to have arisen from an ancestral promoter-replicator sequence which was at the origin of the nonanucleotide promoters present in the mitochondrial genomes of several yeasts. Common sequence patterns indicate that the AT spacers so formed gave rise to the *var1* gene (by linking and phasing of short ORFs), to the DNA stretches corresponding to the untranslated mRNA sequences and to the central stretches of *ori* sequences from *S. cerevisiae*.

Correspondence to: Dr. G. Bernardi, Laboratoire de Génétique Moléculaire, Institut Jacques Monod, Tour 43, 2, Place Jussieu, 75005 Paris (France) Tel. (1)43 29 58 24; (1)43 36 25 25, ext. 41.01.

* This paper is dedicated to the memory of our colleague Renzo Marotta (†1986).

Abbreviations: aa, amino acid(s); A/T, G/C stand for A and/or T, G and/or C, respectively; bp, base pair(s); CRF, closed reading frame; kb, 1000 bp; nt, nucleotide(s); ORF, open reading frame; *ori*, origin of DNA replication; *R* ratio, AT + TA/AA + TT molar ratio; UTS, untranslated sequences; ρ, buoyant density.

INTRODUCTION

Investigations started 20 years ago in our laboratory very soon led to the demonstration (Bernardi et al., 1968) that the mitochondrial DNA from a cytoplasmic petite mutant of *S. cerevisiae*, DM1, had a GC content of only 4%, and a buoyant density, 1.672 g/cm^3, strikingly lower than that of mitochondrial DNA from wild-type cells, viz. 1.685 g/cm^3 (Tewari et al., 1965; Corneo et al., 1966). These results not only unambiguously substantiated the observation (Mounolou et al., 1966) that mitochondrial DNA from petite mutants could be different in buoyant density from that of wild-type cells, but also provided the first evidence that the difference was due to a different base composition and not to the presence of odd bases or to post-synthetic modifications. They also indicated the presence in the mitochondrial genome of stretches almost exclusively made of A and T, raising the problem of their origin and their biological significance. The fact that the buoyant density of DM1 DNA was lower than that of poly(AT : AT) (ρ = 1.678 g/cm^3; Schildkraut et al., 1962) was explained as due to the presence of lighter poly(A : T) stretches (ρ = 1.647 g/cm^3; Szybalski, 1968).

Subsequent work (Bernardi et al., 1970; Bernardi and Timasheff, 1970) showed that the mitochondrial genome of wild-type yeast cells, which is only 17.5% in GC, was largely made up of DNA stretches almost exclusively formed by alternating and non-alternating A : T sequences, the former being predominant (Ehrlich et al., 1972), and proved that these stretches were responsible for a number of 'anomalous' physical properties of yeast mitochondrial DNA. Micrococcal nuclease degradation (Prunell and Bernardi, 1974) demonstrated that the AT spacers had a GC content lower than 5% and represented about 50% of the mitochondrial genome, and that 10% of the genome consisted of short stretches 65% in GC. Degradations with *Hae*III and *Hpa*II provided further information on the structure and the number of such short stretches, the GC clusters (Prunell et al., 1977; Prunell and Bernardi, 1977).

Although a considerable amount of precise information on the long AT spacers and the short GC clusters had been collected in pre-sequencing days, obviously more details became available after primary structures began to be known (Van Kreijl and Bos, 1977; Cosson and Tzagoloff, 1979; Gaillard and Bernardi, 1979). In particular, the internal repetitiveness of the AT spacers was confirmed (Bernardi and Bernardi, 1980), and our original suggestion (see Piperno et al., 1972) that direct repeats in the AT spacers (as well as in the GC clusters; Prunell et al., 1977) were involved in the excision of petite genomes could receive direct experimental support (Gaillard et al., 1980; de Zamaroczy et al., 1983).

Questions were, then, focused on (i) the evolutionary origin and (ii) the possible functional role of the AT spacers, of GC clusters and of intergenic sequences in general. As far as the first point was concerned, it was proposed (Bernardi, 1982) that AT spacers originated by a slippage mechanism, followed by duplication events, from the recently discovered *ori* sequences (de Zamaroczy et al., 1979; 1981; Goursot et al., 1980; Blanc and Dujon, 1980); likewise, GC clusters were suggested to have originated from the GC clusters present in the *ori* sequences. More recent work (de Zamaroczy and Bernardi, 1986a) indicated that the GC clusters of the mitochondrial genome indeed arose from an initial GC cluster present in a 'primitive' *ori* sequence; moreover, it suggested that the starting sequences for AT spacers could be two short sequences flanking the initial GC cluster of the 'primitive' *ori* sequence, namely the *r* sequence (which comprises the nonanucleotide where RNA primers for nascent DNA chains are initiated), and the *r** sequence fulfilling a similar function for the replication of the other DNA strand (Baldacci et al., 1984). As for the second point, arguments were provided (Bernardi, 1982; 1983; de Zamaroczy and Bernardi, 1985; 1986a) in favor of the idea that intergenic sequences did play a functional role.

In the present work, we have analyzed in detail all the available intergenic AT-spacer sequences, as well as the central *I* stretches of *ori* sequences, the sequences corresponding to the untranslated portions of mature RNAs, the five intergenic ORFs and the *var1* gene; moreover, we have critically re-examined the evolutionary origin and the mechanism of formation of the AT spacers and of the *var1* gene, as well as the possible function(s) of the former. Our conclusions on the evolutionary issues have been greatly aided by parallel investigations, also reported here, on the mitochondrial genome of *T. glabrata*.

MATERIALS AND METHODS

We have recently compiled and collated all sequence data available for the mitochondrial genome of *S. cerevisiae* (de Zamaroczy and Bernardi, 1986b). Using this data base in the present work, all sequenced intergenic AT spacers (from the non-transcribed strand) have been assembled end-to-end into a 'compiled' AT spacer, in which appropriate interruption signals were inserted to replace 'gaps', genes, *ori* sequences, ORFs and GC clusters; this was done to avoid artificial 'joint sequences'. Such a 'compiled' AT spacer, which has a total size of 32000 nt, and corresponds to 80% of the total AT spacer, has been analyzed in its 'sequence pattern', by assessing the frequencies of several series of isostichs 2–14 nt long and comparing them with statistical expectations.

Moreover, the distribution of A/T isostichs from the compiled AT spacer of *S. cerevisiae* was compared with (i) that obtained from the 'compiled' AT spacer of *T. glabrata* (based on sequence data of Clark-Walker et al., 1985); and (ii) that of the *var1* gene from *S. cerevisiae* and *T. glabrata*.

The computer programs used were either available at the CITI2 Center, in Paris, or specially designed by Claude Mugnier for the random generation of sequences ('random' and 'simulated' spacers) and for the study of the A/T isostich distribution. Additional details on methods and computing are given in the text and in the figure legends.

RESULTS

(a) The AT spacers of the mitochondrial genome of *Saccharomyces cerevisiae*

The 'long' mitochondrial genome of *S. cerevisiae* has a size of 85 kb and is fully sequenced except for 13 sequence 'gaps', which only represent 8% of the total size (de Zamaroczy and Bernardi, 1985; 1986b). As shown in Table I, intergenic sequences

TABLE I

The 'long' mitochondrial genome of *Saccharomyces cerevisiae*[a]

Elements	Genomic content[a] (%)
Genes (exons)[b]	16
Introns	21
ORFs[b]	(15)
CRFs	(6)
Intergenic sequences[c]	63
ORFs[b]	(5)
ori sequences	(3)
GC clusters[d]	(8)
AT spacers[e]	(47)

[a] The size of the 'long' genome is 85 kb (taken here as 100%). This includes 13 sequence gaps, corresponding to 8% of the genome (de Zamaroczy and Bernardi, 1985; 1986b).

[b] Actually and potentially coding sequences represent 36% of the genome; at least 64% of the genome are, therefore, formed by non-coding sequences.

[c] Including gap sequences except for gap 13, located in the CRF of *cob* intron bI2.

[d] Excluding those (1–2%) present in *rRNA* and *var1* genes, intronic CRFs, intergenic ORFs and *ori* sequences (de Zamaroczy and Bernardi, 1986a).

[e] Including untranslated sequences. The sequenced AT spacers, correspond to 38% of the genome, and to 80% of all AT spacers.

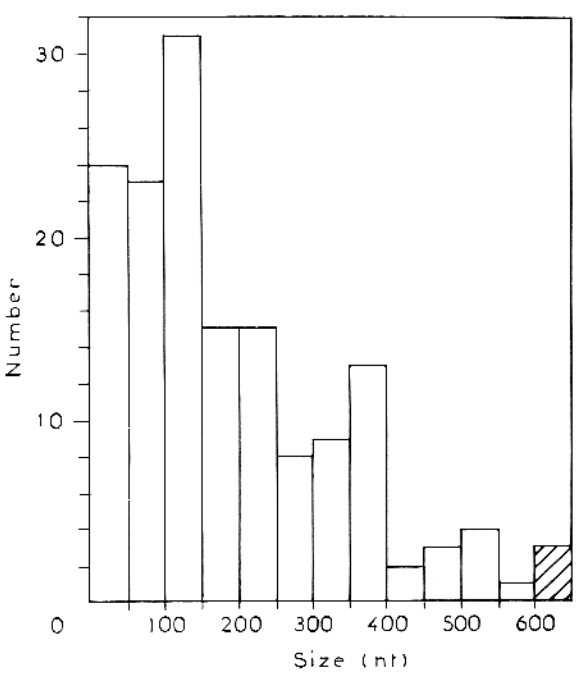

Fig. 1. Size distribution of AT spacers from the mitochondrial genome of *S. cerevisiae* (see MATERIALS AND METHODS), as separated by genes, intergenic ORFs, *ori* sequences and GC clusters; spacers containing gaps were neglected, thus reducing the number of AT spacers taken into consideration to 151. The hatched bar corresponds to sizes comprised between 600 and 800 nt.

represent 63% of the genome (assuming that they also form the sequences corresponding to the gaps, except for gap 13 which is located in the CRF of *cob* intron bI2). The G + C content of intergenic sequences is equal to 14%, a value lower than that of the whole genome, 17.5% (Bernardi et al., 1970; de Zamaroczy and Bernardi, 1985).

Intergenic sequences are formed by 170–200 AT spacers, which correspond to 47% of the genome and are separated from each other by GC clusters, ORFs, *ori* sequences, as well as by genes (primary structure data are presented in de Zamaroczy and Bernardi, 1986b). AT spacers have an average size of 190 bp, the highest frequencies being in the 100–150-bp range (Fig. 1). They are formed by short (20–30 nt on the average), pure A/T stretches separated by C/G mono- to trinucleotides, which account for their 5% GC level.

(b) The sequence pattern of the AT spacers of *Saccharomyces cerevisiae*

(1) The base composition of the 'compiled' AT spacer is given in Table II, and is characterized by essentially equimolar amounts of A and T, as well as of G and C. This base composition is practically the same in six 5000-nt non-overlapping stretches, whereas variations become apparent and base ratios may change at the 2000-nt size level.

(2) The dinucleotide analysis of the 'compiled' AT spacer (Table II) indicates unit molar ratios for all complementary dinucleotides. While this result is compatible with statistical expectations, because of the equimolarities of A and T, and of G and C, respectively, other molar ratios are not. Indeed, AT = TA ≠ AA = TT; CC = GG ≠ GC = CG; AC = TG = CA = GT ≠ AG = TC = GA = CT. If

TABLE II

Base composition, dinucleotide frequencies, and frequency differences of the 'compiled' AT spacers (non-transcribed strand) from the mitochondrial genomes of *Saccharomyces cerevisiae* and *Torulopsis glabrata*

Yeast strain	Base composition[a]			
	A	T	G	C
S. cerevisiae	48.0	47.0	2.4	2.6
T. glabrata	50.8	46.1	1.8	1.3

	Dinucleotide frequencies[a]				
		A	T	G	C
S. cerevisiae	A	18.54	27.26	1.23	0.95
	T	27.45	17.35	0.83	1.29
	G	1.17	0.93	0.31	0.04
	C	0.86	1.33	0.08	0.38
T. glabrata	A	18.32	30.45	1.16	0.68
	T	31.61	13.48	0.61	0.55
	G	0.55	1.16	0.06	0.03
	C	0.49	0.81	0.03	0

	Frequency differences[b]							
	AA	TT	AT	TA	GG	CC	AC	CA
S. cerevisiae	−19.5	−21.5	+20.8	+21.7	+438	+462	−23.9	−31.1
T. glabrata	−29.0	−36.6	+30.0	+35.0	+85.2	−100	+3.0	−25.8

[a] Base compositions and frequencies are given as mol%.

[b] Frequency differences are given as 100 × (found−expected)/(expected), where 'found' refers to the frequencies in the 'compiled' AT spacer and 'expected' to the statistical frequencies.

dinucleotide frequencies for the 'compiled' AT spacer and for 5000-nt stretches are compared with the statistical expectations, some characteristic differences are found (Table II). The typical features are 21% excesses for AT and TA, and 20% deficits for AA and TT; the molar ratio AT + TA/AA + TT, *R*, is equal to 1.5. Other differences concern GG and CC, which show approx. 450% excess, and AC and CA, which show 24–31% deficit. Slight fluctuations of these frequency differences are found in 5000-nt and, more so, in 2000-nt stretches, but the major trends just described may still be seen in stretches as short as 200 nt.

(3) The frequencies of trinucleotides in the 'compiled' AT spacer are shown in Table III. The differences between the trinucleotide frequencies experimentally found in the 'compiled' AT spacer and the statistical frequencies are shown in Tables III and IV. The experimental frequencies of trinucleotides formed by A/T only (which statistically have the same expected frequencies) show 50% excess for ATA and TAT and 40% deficit for AAA and TTT; the other trinucleotides, TTA, ATT, AAT and TAA exhibit frequencies which are comparable with the statistical ones (Table III). In the case of trinucleotides formed by G/C only, GGG and CCC are present in very large excess, whereas CGC and GCG are close to the statistical expectations. Among the other trinucleotides, those comprising two C or two G are in large excess over statistical expectations; in contrast, those comprising one C/G are close to statistical expectations, with some exceptions (Table IV).

(4) The analysis of higher isostichs has been limited to A/T sequences up to 14 nt because these sequences carry the basic features of AT spacers (see above) and because the increase in the number and the decrease in frequency of possible isostichs (same-size sequences) in a higher size range set an upper limit to our analysis.

The analysis of 4- to 14-nt long A/T sequences has shown that isostichs cover a range of frequencies, whereas those from the 'random' AT spacer (namely in a randomly generated sequence having the same base composition and size) are close in frequency. This range increases with increasing isostich size.

TABLE III

Trinucleotide frequencies (**A**) in the mitochondrial genome of *Saccharomyces cerevisiae* and frequency differences of A/T trinucleotides (**B**) of the 'compiled' AT spacers of *S. cerevisiae* and *Torulopsis glabrata* [a]

(**A**)	AAA 6.56	ATA 16.60	AGA 0.56	ACA 0.34
	AAT 10.88	ATT 9.66	AGT 0.47	ACT 0.43
	AAG 0.74	ATG 0.49	AGG 0.19	ACG 0.03
	AAC 0.32	ATC 0.56	AGC 0.02	ACC 0.15
	TAA 11.16	TTA 10.01	TGA 0.42	TCA 0.44
	TAT 15.44	TTT 6.46	TGT 0.33	TCT 0.63
	TAG 0.34	TTG 0.27	TGG 0.07	TCG 0.04
	TAC 0.53	TTC 0.59	TGC 0.02	TCC 0.18
	GAA 0.47	GTA 0.49	GGA 0.16	GCA 0.02
	GAT 0.53	GTT 0.37	GGT 0.11	GCT 0.02
	GAG 0.13	GTG 0.04	GGG 0.04	GCG 0
	GAC 0.04	GTC 0.04	GGC <0.01	GCC 0.01
	CAA 0.34	CTA 0.37	CGA 0.03	CCA 0.08
	CAT 0.45	CTT 0.82	CGT 0.04	CCT 0.24
	CAG 0.02	CTG 0.04	CGG 0.01	CCG 0.01
	CAC 0.06	CTC 0.11	CGC 0	CCC 0.04

(**B**)	AAA	TTT	ATA	TAT	ATT	TTA	TAA	AAT
S. cerevisiae	−40.7	−37.8	+53.4	+45.7	−8.9	−5.6	+3.1	+0.6
T. glabrata	−57.5	−60.2	+78.6	+70.7	−18.1	−15.9	+4.7	−2.2

[a] See footnotes a and b of Table II.

TABLE IV

Frequency differences of trinucleotides for the 'compiled' AT spacer of *Saccharomyces cerevisiae*

Trinucleotides	Statistical frequency [a]	Frequency differences [b,c]	
A/T trinucleotides	10.7	± 1	(see Table III)
TTC CTT TCT			± 14
AAC CAA ACA	0.7	± 15	− 52
ATC CTA TAC CAT ACT TCA			− 20 − 40 − 24
AAG GAA AGA			+ 61 + 2 + 22
TTG GTT TGT	0.45	± 12	− 39 − 16 − 25
ATG GTA TAG GAT AGT TGA			± 15
CCT TCC CTC	0.04	± 25	+ 500 + 350 + 175
CCA ACC CAC			+ 100 + 275 + 50
ACG GCA CAG GAC CGA AGC	0.03	± 70	± 30
TCG GCT CTG GTC CGT TGC			± 30
GGA AGG GAG	0.02	± 70	+ 700 + 850 + 550
GGT TGG GTG			+ 450 + 250 + 100
GGG	0.001	+ 100	+ 3900
CCC	0.003	− 100	+ 1233
CCG GCC CGC	0.002	± 100	+ 400 [d]
GGC CGG GCG	0.001	± 200	+ 500 [d]

[a] As calculated from the base composition.

[b] The first column presents the frequency differences for trinucleotides found in a 'random' AT spacer, namely in a randomly generated sequence having the same base composition and size, and carrying the same interruption signals, as the 'compiled' AT spacer (in fact, the signals do not appreciably influence the statistical values) compared to the statistical frequencies, calculated from the base composition. The second column presents differences between the frequencies experimentally found for trinucleotides in the 'compiled' AT spacer (see Table III) and the frequencies calculated from the base composition. The comparison of the values from the second column with those from the first column allows an estimation of the significance of the former. Frequency differences are given as indicated in Table II, footnote [b].

[c] Underlined values refer to the underlined trinucleotides. Averages are given for close values.

[d] CGC and GCG frequencies are very low and frequency differences are not significant.

For instance, in the case of hexanucleotides (Table V), frequencies range from 1190 to 65 copies, instead of being equal to 350 (± 13%), as is the case for the 'random' AT spacer (Table V).

Since the same basic rules apply to all isostich classes, these will be presented for just two of them, tetra- and hexanucleotides. Table V and Fig. 2a show that: (i) the most frequent tetra- and hexanucleotides are those made of (or enriched in) alternating AT only, whereas the least frequent ones are those enriched in both non-alternating A and non-alternating T; (ii) the frequency decrease is related to the decrease in the R ratio of isostichs; this R ratio defines classes of tetra- and hexanucleotides showing close frequencies (generally within the range found for statistical sequences); at the hexanucleotide (but not at the tetranucleotide) level, such classes may comprise sub-classes; moreover, some hexanucleotides (AATAAT, TAATAA, TTTATT, AATAAA, AAAAAA and their complementary sequences) show the frequencies expected for classes defined by immediately higher R ratios; (iii) the average frequencies of hexanucleotides from different classes exhibit an exponential relationship with the cor-

TABLE V

Frequency (copy number) distribution of A/T tetra- and hexanucleotides from intergenic sequences. Data from both *S. cerevisiae* and *T. glabrata* are presented[a]

	R Ratio (A)	*S. cerevisiae* (B)	*S. cerevisiae* (C)	*T. glabrata* (D)	*T. glabrata* (E)
4-mers					
TATA	3:0	3118	3099	396	387
AATA	2:1	2534	2173	239	149
ATAA	2:1	2506	2198	235	154
TAAT	2:1	2426	2073	254	171
TTTA	1:2	1306	1201	90	94
TAAA	1:2	1301	1187	104	81
TTAA	1:2	1290	1102	122	107
AAAA	0:3	893	884	58	29
6-mers					
ATATAT	5:0	1190	1186	181	194
ATAATA	4:1	969	746	115	45
AATAAT	3:2	762	606	74	31
TAATAA	3:2	787	536	69	26
ATATAA	4:1	700	597	87	51
AATATA	4:1	702	578	83	57
TAATAT	4:1	687	579	91	65
TATAAT	4:1	639	593	76	54
ATAAAT	3:2	550	500	34	38
ATTAAT	3:2	564	484	58	61*
TAAATA	3:2	554	486	41	31
TTAATA	3:2	529	370	42	44
TATTAA	3:2	472	417	42	32
TTTATT	2:3	440	421	21	33
TATAAA	3:2	454	395	39	33
AAATAT	3:2	418	418	36	25
AATATT	3:2	418	417	45	33
AATAAA	2:3	441	378	26	17
TTAATT	2:3	331	316	42	32*
TTTTAT	2:3	334	281	12	26
ATAAAA	2:3	315	283	26	15
TAATTT	2:3	242	239	21	16
TTTAAT	2:3	291	187	26	15
ATTTTA	2:3	239	184	15	21
ATTTAA	2:3	229	181	29	16
AATTTA	2:3	215	195	23	22
TAAAAA	1:4	231	145	13	3
TTTTTA	1:4	188	177	3	10
AAAAAA	0:5	176	170	9	7
TTAAAA	1:4	109	101	6	3
TTTTAA	1:4	118	76	7	5
TTTAAA	1:4	73	65	12	4

[a] In the 'random' AT spacer (related to that of *S. cerevisiae*), copy numbers for tetra- and hexanucleotides are 1600 (±6%) and 350 (±13%), respectively. In this analysis, intergenic sequences from *S. cerevisiae* do not comprise intergenic ORFs. Column (A): (AT + TA)/(AA + TT) molar ratio. Columns (B), (C), (D), and (E): copy numbers of tetra- and hexanucleotides (B and D) and their complementary sequences (C and E). Ordering of isostichs is according to the copy numbers of the sums of complementary sequences from *S. cerevisiae*. Hexanucleotides showing higher copy number than expected on the basis of their ratio are underlined (except for AATAAA in the case of *T. glabrata*). Asterisks indicate hexanucleotides which show the copy number expected from their ratio in the case of *S. cerevisiae*, but a higher copy number in the case of *T. glabrata*.

responding *R* ratios (see Fig. 4). (It should be mentioned here that a similar distribution of A/T hexanucleotides is obtained for two 'simulated' AT spacers randomly generated from pools of AA, TT, AT and TA in appropriate proportions. The DISCUSSION section presents detailed comments on these results shown in Figs. 3 and 4.)

From a 10-nt size level on, not all possible isostichs are present. For instance, in the case of 14-mers, only 48% of the possible 16 384 isostichs are found, 73% of them just once or twice. Higher frequencies are exceedingly rare and concern 758 14-mer sequences which are present more than four times (only 148 of them are present more than ten times).

These most represented sequences generally comprise two directly repeated hexanucleotides, which are among the most frequent ones. In other words, higher-size oligodeoxynucleotides largely consist, as expected, of lower-size oligodeoxynucleotides belonging to the most frequent classes. This leads to an abundance (compared to the 'random' AT spacer) of overlapping repeats and also of shared repeats (i.e., repeats shared by higher-size oligodeoxynucleotides). For instance, the most frequent 14-mer sequences consist of internally repeated AT (or TA, if the sequence is read in the other frame), AAT (or ATA, or TAA), AATT (or ATTA, or TTAA, or TAAT) and TTTA (or TATT, or ATTT, or TTAT); these four 14-mer sets comprise 108, 61, 36 and 31 copies, respectively; they generally belong to much longer physical sequences and are, therefore, counted several times over, which explains their high frequencies.

The extreme cases of alternating AT and non-alternating A or T sequences were investigated at different size levels (Fig. 5). The ratio of the former sequences to statistical expectation increases exponentially with sequence size; in contrast, non-alternating sequences show essentially a constant

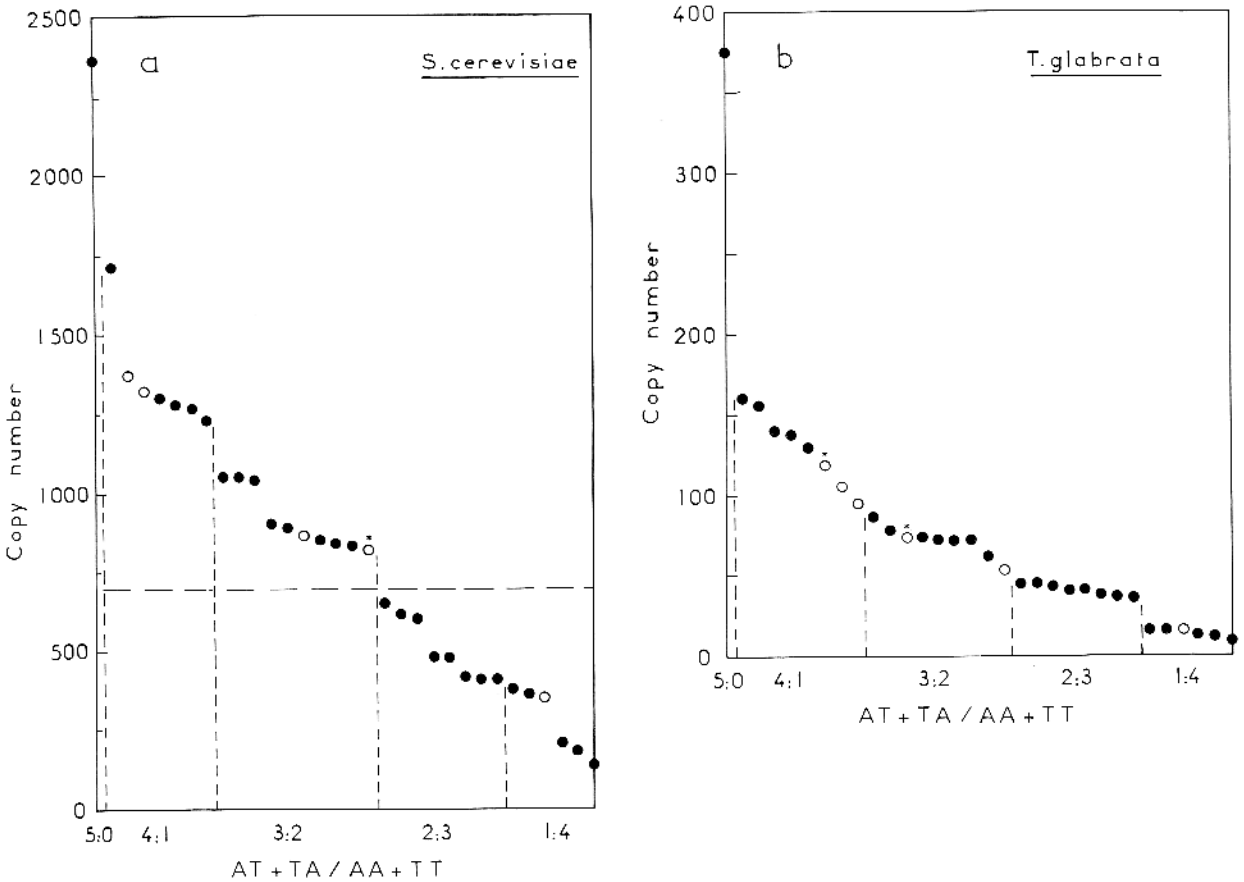

Fig. 2. Copy numbers of A/T hexanucleotides (see Table V; data points, closed and open circles correspond to average values of pairs of complementary sequences) from the 'compiled' AT spacers for *S. cerevisiae* (Panel **a**) and *T. glabrata* (Panel **b**) are shown in order of decreasing frequency. The corresponding *R* ratios are indicated on the abscissa. Open circles concern hexanucleotides which are more frequent than expected from their ratio (see Table V). These points correspond to the same hexanucleotides in *S. cerevisiae* and *T. glabrata*, except for those marked by a tiny asterisk. The horizontal broken line corresponds to the distribution expected for a 'random' sequence having the base composition and size of the 'compiled' AT spacer from *S. cerevisiae*.

ratio with statistical expectations, both distributions being essentially exponential. The upper limit of alternating AT sequences is twice as high (24 nt) as that of non-alternating A or T sequences (12 nt).

As expected, an analysis of pairs of non-overlapping direct repeats (and complementary inverted repeats; not shown) indicates that their distribution reaches much higher sizes in the 'compiled' AT spacer than in the 'random' AT spacer, the latter exhibiting an upper limit of 23 nt. Only 12 pairs of direct repeats longer than 30 nt (the longest one being 41 nt) have been found (Table VI); they never contain G/C and they all comprise 3–7-nt-long internal repeats, which are often contiguous. Moreover, two pairs correspond to contiguous repeats which are only made up of short internal repeats (Table VI).

(c) The AT spacers of the mitochondrial genome of *Torulopsis glabrata*

The 31 intergenic AT spacers of the mitochondrial genome (19 kb) of *T. glabrata* (Clark-Walker et al., 1985) do not contain any GC cluster and belong to two classes: ten of them are 240-nt long on the average and 21 are 20-nt long on the average and located between tRNA genes.

Tables II and III show that the base composition of the 3117-nt-long 'compiled' AT spacer (assembled end-to-end from the non-transcribed strand) is similar to that of *S. cerevisiae*. (A is, however, slightly more abundant than T; the GC level, 3.1%, is lower; and CC and GG are not predominant among the few C/G dinucleotides.) The similarity extends to the A/T di- and trinucleotides which show the typical

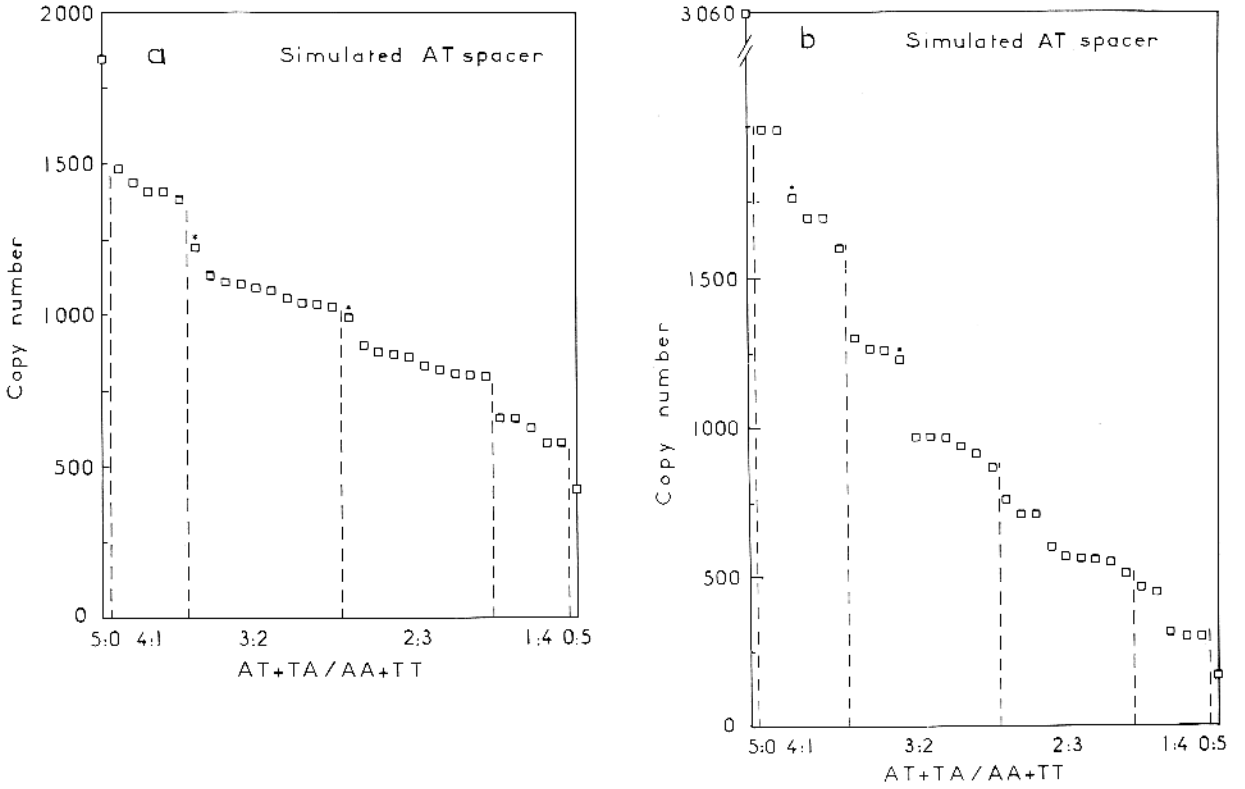

Fig. 3. Copy number of A/T hexanucleotides (average values of pairs of complementary sequences) from the 'simulated' AT spacer constructed by random assortment of A/T dinucleotides and showing final *R* ratios of 1.3 (a) and 1.7 (b) (see DISCUSSION, section **b**). Hexanucleotides are shown in order of decreasing frequency. The corresponding *R* ratios are indicated on the abscissa. Tiny asterisks above squares indicate hexanucleotides which are more frequent than expected from their ratio.

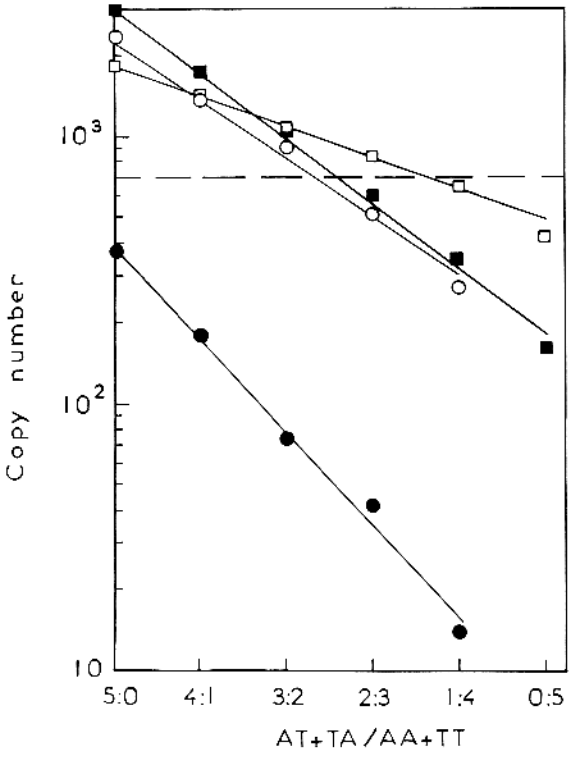

Fig. 4. Average copy numbers of A/T hexanucleotides belonging to the same class (see Table V and Figs. 2 and 3) are plotted against their *R* ratios. Open and solid circles refer to *S. cerevisiae* and *T. glabrata* data, respectively; open and solid squares to the 'simulated' AT spacers having *R* ratios of 1.3 and 1.7, respectively. The horizontal broken line corresponds to the distribution expected for a 'random' sequence having the base composition and size of the 'compiled' AT spacer from *S. cerevisiae*.

TABLE VI

Primary structure of non-overlapping direct repeats longer than 30 nt in the 'compiled' AT spacer and in the *var1* gene of *Saccharomyces cerevisiae*[a]

$(AATATTT)_6$
$T(AATT)_7$ $(AAT)_2$ AAA
$(TA)_3$ TTTAATATTATTTTT$(AT)_2$ AATTTTATATTA
$(TAAAAA)_6$[b]
$ATT(ATTT)_8$[c]
$ATTT(AT)_2$ $A(AT)_3$ $(TA)_2$ $AA(TAA)_4$
AATAT(AAATAT)$_2$ AATATATTTTAATAT[b]
$(ATATT)_6$
$(TTTTA)_6$[b,d]
$(TAA)_{10}$[d,e]
$(AT)_5$ $TA(TAA)_6$
TTATT$(AT)_3$ $(TAA)_2$ $(TA)_4$ TTATT

[a] Sequences in parentheses and underlined sequences show short repeated sequences within the long repeats.
[b] In this case the pair of long repeats are contiguous.
[c] This repeat exists in three copies.
[d] Long repeats only present in some strains (polymorphism).
[e] This only pair of long repeats from the *var1* gene is located at *b1* and *b2* inserts.

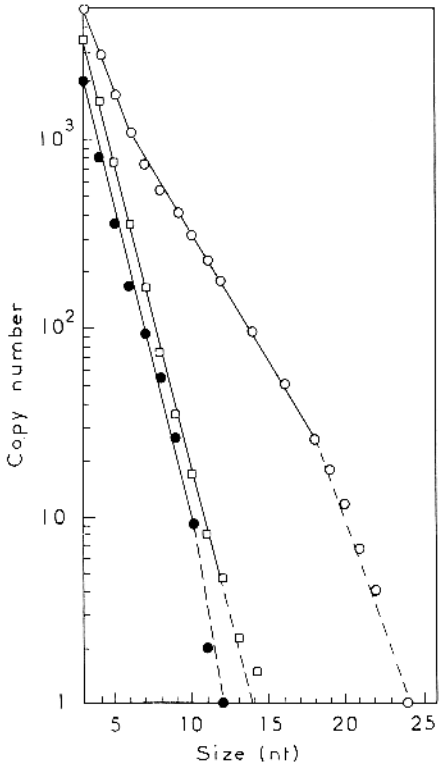

Fig. 5. Size distributions of alternating AT (open circles) and non-alternating A or T oligodeoxynucleotides (solid circles) in the 'compiled' AT spacer from *S. cerevisiae* and in an AT spacer constructed by random assortment of A, T, G and C (open squares). Size distributions of alternating sequences were the same for sequences starting with A or T; likewise, size distributions of oligo(A)'s and oligo(T)'s were the same. Frequencies shown correspond to the average of the two classes, or of the four classes in the case of the 'random' sequence. The broken line corresponds to a region of data scatter, because of low frequencies.

frequency deviations found in *S. cerevisiae*; the deviations are, in fact, even stronger, and the *R* ratio is 2.0. The non-stoichiometry of A and T and the difference in *R* ratio are, in all likelihood, due to the small size of the AT spacer from *T. glabrata*, since similar differences have been found in stretches of comparable size from the AT spacer of *S. cerevisiae*.

The distribution of A/T tetra- and hexanucleotides (Table V) shows, as in the case of *S. cerevisiae*, an exponential decrease of frequencies with decreasing *R* ratio (Fig. 2b), the slope in a plot against *R* being higher than that of *S. cerevisiae* (Fig. 4). Minor differences concern the frequency order of oligodeoxynucleotides belonging to the same class (as defined by the same ratio *R*). Moreover, four out of the five hexanucleotides (AATAAT, TAATAA, TTTATT and AAAAAA), which show strong frequency excesses in *S. cerevisiae*, as well as two additional hexanucleotides, ATTAAT and TTAATT, show the frequencies expected for classes defined by immediately higher *R* ratios. No sub-classes are detectable in the classes of *T. glabrata*, probably because of the small size of the AT spacer.

(d) The *var1* genes of the mitochondrial genomes of *Saccharomyces cerevisiae* and *Torulopsis glabrata*

If its GC cluster is neglected, the composition of the *var1* gene [42.0], of *S. cerevisiae* (Hudspeth et al., 1982) is characterized by a higher G + C content and an asymmetry of A and T (Table VII) compared to the 'compiled' AT spacer (or to flanking AT spacers having the same size as the gene, and likewise considered without their GC clusters; not shown). The frequency differences of A/T di- and trinucleotides and of CC, GG are similar to those of the AT spacer (and of the flanking sequences) except that TAT shows no excess and that AAT and TAA show a large excess. Other similarities with the AT spacer concern (i) the existence of a strain polymorphism (de Zamaroczy and Bernardi, 1986b), consisting in the lower or higher repetition of a short A/T oligonucleotide (two runs of AAT repeats, in the case of *var1* gene; Hudspeth et al., 1982); (ii) the presence within the gene of one or, occasionally, two (Hudspeth et al., 1984) GC cluster(s) belonging to a family also represented elsewhere in the genome (de Zamaroczy and Bernardi, 1986a); and (iii) the distribution of A/T tetra- and hexanucleotides (see below).

The *var1* gene of *T. glabrata* (Table VII) is almost identical in base composition, and in the frequency deviations of A/T di- and trinucleotides and of CC and GG, to that from *S. cerevisiae*, with which it shows 85.3% nucleotide sequence homology (Ainley et al., 1985); the *T. glabrata* gene is, however, shorter by 150 nt, because of several additions/deletions and the absence of GC clusters.

The distribution of A/T tetra- and hexanucleotides (Table VIII) follows the same rules found for the intergenic AT spacers. The most frequent sequences in each complementary pair are practically the same for the two *var1* genes and the corresponding intergenic AT spacers.

The separation among hexanucleotide classes as defined by the *R* ratio is, however, not evident anymore, in all likelihood because of both the narrower frequency range covered and of two common specific sequence features, which distinguish them from the intergenic AT spacers: (i) the most frequent oligodeoxynucleotides from the latter, the alternating AT sequences are much less frequent in the *var1* genes; moreover, other hexanucleotides, which comprise those having a 4 : 1 *R* ratio (except for ATAATA), as well as TATAAA and AAATAT, also show in both species much lower frequencies than expected on the basis of the corresponding *R* ratio; all these hexanucleotides share an ATAT and/or TATA sequence; (ii) complementary sequences do not have comparable frequencies; in spite of sharing the same ratios, they show frequencies which differ by a factor of 2–3, the higher values generally corresponding to the hexanucleotides enriched in A.

Finally, it should be noted that the frequencies of three hexanucleotide sequences show different frequencies in the *var1* genes of *S. cerevisiae* and *T. glabrata*; the higher frequency of AATAAT, ATAATA and TAATAA in the former is due to the presence of *b1* and *b2* inserts (AAT runs) in the [42.0] allele compared to the [40.0] allele.

TABLE VII

Base composition and frequency differences of di- and trinucleotides for the *var1* gene of *Saccharomyces cerevisiae* and *Torulopsis glabrata*[a]

	A	T	G	C	Size (nt)			
S. cerevisiae	52.2	40.1	4.3	3.4	1169			
T. glabrata	51.8	41.0	4.9	2.3	1017			
	AA	TT	AT	TA	GG	CC	AC	CA
S. cerevisiae	−12.7	−27.6	+18.6	+21.9	+ 216	+258	−13.5	−42.1
T. glabrata	−11.3	−22.1	+14.9	+19.6	+ 192	+467	−16.1	−41.3
	AAA	TTT	ATA	TAT	ATT	TTA	AAT	TAA
S. cerevisiae	−47.2	−58.5	+38.5	+ 4.8	− 7.1	− 3.6	+32.1	+39.5
T. glabrata	−42.5	−63.9	+19.1	+ 3.4	+10.3	+16.1	+30.0	+33.6

[a] See footnotes a and b of Table II. The GC cluster of the *var1* gene from *S. cerevisiae* was neglected in this analysis.

TABLE VIII

Frequency (copy number) distribution of A/T tetra- and hexanucleotides (partial list) from the *var1* gene (non-transcribed strand) of *Saccharomyces cerevisiae* and *Torulopsis glabrata*[a]

	Ratio *R*	*S. cerevisiae*		*T. glabrata*	
	(A)	(B)	(C)	(D)	(E)
TAAT	2:1	108	65	88	78
ATAA	2:1	117	52	83	51
AATA	2:1	121	40	98	40
ATAT!	3:0	53	43	44	33
TTAA	1:2	48	35	55	39
AAAT	1:2	57	23	51	20
TAAA	1:2	60	19	54	15
AAAA	0:3	22	7	23	5
AATAAT	3:2	64	16	39	20
ATAATA	4:1	65	13	42	13
TAATAA	3:2	66	11	44	12
TAATAT!	4:1	21	13	22	14
TAAATA	3:2	26	8	20	5
TATTAA	3:2	22	11	23	8
ATATTA!	4:1	19	13	18	13
TTAATA	3:2	17	14	24	14
AATATT	3:2	20	10	20	8
ATAAAT	3:2	23	6	17	6
ATTAAT	3:2	17	8	26	8
ATATAA!	4:1	17	6	12	7
TTAAAT	2:3	17	6	16	4
AATATA!	4:1	16	6	10	8
TATAAA!	3:2	14	8	8	8
AAATAA	2:3	16	5	14	6
ATTAAA	2:3	16	5	16	3
AAATTA	2:3	17	2	21	2
AATAAA	2:3	17	2	18	0
TAAATT	2:3	13	4	15	6
AAATAT!	3:2	13	3	11	2
AATTAA	2:3	10	3	16	3
ATATAT!	5:0	4	4	1	3
AAAAAA	0:5	4	0	3	0

[a] Symbol ! indicates isostichs which show much lower copy number than expected on the basis of their ratio. Columns (A)–(E): see corresponding footnote to Table V.

(e) Other intergenic sequences of *Saccharomyces cerevisiae*

(1) The l *stretches of* ori *sequences*

These central 200-nt A + T-rich segments of *ori* sequences are characterized by a very strong conservation in both primary structure and length, by a large predominance of A↔T transversions over transitions and by an asymmetrical distribution of A and T on the two strands (de Zamaroczy et al., 1984). Their base composition is different from that of the 'compiled' AT spacer, but di- and trinucleotide frequencies show the typical deviations from statistical expectations presented by the 'compiled' AT spacer; other deviations concern the trinucleotides, ATT, TTA, AAT and TAA, which show no significant deviations in the 'compiled' AT spacer (Table IX). Interestingly, the 'majority sequences' of *l* stretches from all *ori* sequences (see Fig. 6 in de Zamaroczy et al., 1984) largely consists of three short, essentially non-overlapping repeated sequences, TTTA, TAATA(T), and ATAT.

(2) The sequences corresponding to the 5′- and 3′-untranslated mature mRNAs (UTS)

Neglecting the GC clusters which are regularly present, these sequences represent a total of 3120 nt (namely 8% of the AT spacer) in the cases in which they are clearly defined. The base composition and the frequency deviations of A/T di- and trinucleotides, and of CC, GG are indistinguishable from those of AT spacers (Table IX). Individual sequences exhibit the same features as the overall sequence (not shown), except for those corresponding to the 3′ UTS of *oxi3* and *oli1*; this might be due to the small sizes, about 80 nt, of these sequences.

(3) The intergenic ORFs

The base composition of intergenic ORFs (see Colin et al., 1985, for the nomenclature used and for references) shows that three of them, 1, 2 and 4, are practically identical and higher in GC level (18.5% vs. 14%) than intergenic sequences. If GC clusters are neglected, the composition of these three ORFs (Table IX) still remains very similar and higher in GC than the AT spacer or the *var1* gene (15.5% vs. 5% and 8%, respectively), whereas the other two become very close in GC level to the 'compiled' AT spacer. The analysis of A/T di- and trinucleotides (Table IX) confirms that ORFs 1, 2 and 4 are similar to each other and largely different from the AT spacer in the extent of the typical deviations (although less so for ORF4). Particularly, in spite of a high GC level, GG and CC do not show the typical deviations present in AT spacers and in the two *var1* genes. Moreover, frequency deviations of other A/T

TABLE IX

Base composition and frequency differences of di- and trinucleotides for the *l* stretch of *ori*5, for the untranslated sequences (UTS) and for the intergenic ORFs of the mitochondrial genome from *Saccharomyces cerevisiae*[a]

	A	T	G	C	Size (nt)			
*ori*5 *l*	59.5	31.5	3.5	5.5	200			
UTS	49.6	45.7	2.6	2.0	3120			
ORF1	43.3	39.5	9.8	7.3	1119			
ORF2	43.4	40.5	8.8	7.3	1268			
ORF4	45.2	41.6	7.0	6.2	1421			
ORF3	46.3	46.8	4.2	2.6	190			
ORF5	45.5	51.5	1.0	2.0	101			
	AA	TT	AT	TA	CC	GG		
*ori*5 *l*	−17.5	−44.4	+ 31.6	+26.2	—	—		
UTS	−18.3	−22.0	+ 21.2	+20.7	+ 400	+187		
ORF1	− 2.7	− 4.5	+ 5.8	+ 4.1	+ 120	+ 30		
ORF2	− 1.6	− 5.5	+ 2.3	+ 6.3	+ 37.5	+ 80		
ORF4	− 5.9	− 8.7	+ 10.1	+ 7.5	+ 100	+ 80		
ORF3	−21.0	−10.5	+ 22.1	+12.0	—	—		
ORF5	−37.2	−32.1	+ 36.8	+32.5	—	—		
	AAA	TTT	ATA	TAT	ATT	TTA	AAT	TAA
*ori*5 *l*	−52.1	−67.7	+ 71.4	+28.8	− 32.2	− 40.7	+35.7	+40.2
UTS	−40.5	−30.0	+ 55.7	+40.7	− 14.5	− 12.0	+ 4.3	+ 6.2
ORF1	0	−29.0	− 5.4	+ 1.5	+ 16.2	+ 13.2	+ 1.4	+ 6.8
ORF2	− 1.2	−22.7	+ 2.6	+ 5.6	+ 9.9	+ 9.9	− 2.6	+ 5.3
ORF4	−13.0	−27.8	+ 4.7	+14.1	+ 11.5	+ 11.5	+ 7.1	+ 8.2
ORF3	−51.5	−27.2	+ 38.0	+20.8	+ 10.9	+ 5.0	+ 6.0	+ 6.0
ORF5	−78.7	−40.9	+107.5	+66.9	− 24.8	− 33.1	− 5.6	+ 3.7

[a] See footnotes a and b of Table II. The GC clusters of the UTSs and ORFs are omitted from this analysis. Untranslated mature mRNA sequences were compiled end-to-end; they comprise those of *oxi2* (5′; Thalenfeld et al., 1983); *15S* (5′; Christiansen and Rabinowitz, 1983); *oxi3* (5′ and 3′; Osinga et al., 1984a,b); *cob* (5′ and 3′; Bonitz et al., 1982; Osinga et al., 1984b); *oli1* (5′ and 3′; Edwards et al., 1983; Osinga et al., 1984b); *var1* (3′; Osinga et al., 1984b). For the nomenclature of intergenic ORFs and the corresponding references, see Colin et al. (1985).

trinucleotides are different from those of the *var1* gene. In contrast, ORFs 3 and 5 show features which are very close to those of the AT spacer.

DISCUSSION

(a) The sequence pattern of the AT spacers of the mitochondrial genomes of *Saccharomyces cerevisiae* and *Torulopsis glabrata*

The experimental approach followed in the present work to characterize the complex primary structure of the AT spacers from the mitochondrial genomes of *S. cerevisiae* and *T. glabrata* has consisted in defining its sequence pattern, namely in assessing the frequencies of its oligodeoxynucleotides and in comparing them with statistical expectations. This approach has shown that the AT spacers of both mitochondrial genomes are characterized by several common features which open the way to an understanding of their mechanism of formation and their evolutionary origin (see the following sections).

(i) In the case of *S. cerevisiae*, the enormous excess of GGG and CCC (as well as of GG and CC) over statistical expectations is in sharp contrast with the statistical frequency observed for most other G/C containing trinucleotides (as well as for CG and GC). This suggests that the former have a different origin compared to the latter. Indeed, the former are likely to be related to GC clusters (as already proposed; de Zamaroczy and Bernardi, 1986a), whereas

the latter might be due to point mutations (and, possibly, also to the presence of G/C in the initial sequence used in the formation of AT spacers; see below). Along the same line, if GG, CC, GGG, and CCC are neglected, the GC level of AT spacers drops to 3.4%, a value close to the 3.1% of the evolutionarily related AT spacers from the mitochondrial genome of *T. glabrata*, which does neither contain GC clusters, nor show the CC,GG frequency excesses.

(ii) The equimolarity of A and T, their extremely high levels (95%), and the typical frequency deviations of A/T di- and trinucleotides of the 'compiled' AT spacer are found down to very low size levels. This provides a strong indication for a common, evolutionary origin for all the constituent segments of the AT spacers.

(iii) If we now consider in more detail the frequencies of oligodeoxynucleotides only formed by A/T in *S. cerevisiae*, a choice justified not simply by their predominance in the AT spacers, but mainly by their common evolutionary origin (which is separated from that of the G/C sequences discussed above), the following main features appear: (a) among dinucleotides, AT and TA show an excess, AA and TT a shortage, compared to statistical expectations; the *R* ratio is 1.5 for the 'compiled' AT spacer; (b) among trinucleotides, the alternating ones show an excess, the non-alternating ones a shortage; all other combinations are close in frequency to statistical expectations; (c) the frequencies of longer A/T isostichs essentially depend upon their *R* ratio; the average frequencies of the classes, as defined by such ratio, show an exponential distribution; (d) above the 10-nt level, an increasing number of possible isostichs do not occur anymore, and the frequencies of the remaining ones drop to very low levels; expectedly, the frequency of 'unique' sequences (only present once in the AT spacers) increases with increasing size; remarkably, 'unique' sequences are formed by a number of internal non-overlapping and overlapping repeats, which are also shared by other unique sequences (in fact, such unique sequences may differ from each other by as little as a single base); (e) if only non-overlapping sequences are considered, all those longer than 30 nt are unique, except for a small set (see Table VI), only representing 3% of the compiled AT spacer. All the above points also basically apply to the AT spacers of *T. glabrata*.

(iv) The frequency excess of some classes of isostichs relative to their frequencies in a 'random' AT spacer obviously points to their repetitiveness. This adds to a general background of sequence repetitiveness in the AT spacer, which is due to its extreme base composition and to the equimolarity of A and T. Indeed, a 'random' AT spacer shows many more repeated sequences than an equal-size random DNA having a less extreme G + C content. The base-composition effect is by far the predominant factor responsible for sequence repetitiveness in the AT spacer. This is clearly demonstrated by the fact that the homology level of the 'compiled' AT spacer, as determined on 1500-nt stretches taken at random in either orientation, is very high, 68–72%, and essentially the same for the 'compiled' AT spacers of *S. cerevisiae* and *T. glabrata* and the for the 'random' AT spacer. This constant level was found by maximizing alignments with a comparable total number of mono- to trinucleotide gaps. Finally, it should be mentioned that the sequence pattern of the AT spacer is reminiscent of that of some complex satellite DNAs, like the 1.715 satellite from the bovine genome (Gaillard et al., 1981).

(b) The mechanism of formation of the AT spacer of the mitochondrial genome of *Saccharomyces cerevisiae* and *Torulopsis glabrata*

The three main features of the AT spacers from both *S. cerevisiae* and *T. glabrata* are the extremely high repetitiveness of short sequences, the characteristic sequence pattern and the dependence of the frequencies of its A/T sequences upon their *R* ratio.

The first two features (found down to very low sizes) indicate that (i) the AT spacers arose by an expansion process, essentially involving duplications (accompanied by inversions, and translocations); (ii) such mechanism originally operated on an initial oligodeoxynucleotide (see below) and subsequently also on the sequences derived from it; and (iii) the expansion process initially took place in a yeast which was the common ancestor of *S. cerevisiae* and *T. glabrata*. In all likelihood, the expansion mechanism mainly involved recombination events, although replication slippage probably also played a role, particularly at the beginning of the process. Again, in all likelihood, the rounds of recombinational events and the point mutations (essentially

bidirectional A↔T transversions and single base A/T deletions/additions, but also mutations leading to the appearance of C and G) which accompanied them account for the absence of long repeats in the AT spacers.

The third feature of the AT spacers, namely the dependence of the frequencies of A/T sequences upon their *R* ratio in the AT spacer, was used as a starting point for further investigation of the mechanism of formation of the latter and to characterize the initial oligodeoxynucleotide. The approach followed was to randomly generate a sequence having the size and composition of the 'compiled' AT spacer from *S. cerevisiae* using pools of the four A/T dinucleotides with a given initial *R* ratio. Two such 'simulated' AT spacers with final *R* values in the neighborhood of that of the AT spacers (1.3 and 1.7, respectively), were analyzed in their oligodeoxynucleotides and shown to exhibit a striking similarity with the 'compiled' AT spacers of both *S. cerevisiae* and *T. glabrata*. Indeed, the distribution of hexanucleotides (Fig. 3) was very similar to that found for the 'compiled' AT spacer from *S. cerevisiae* (Fig. 2). The similarity also concerned (i) the presence of classes (and subclasses in the 'simulated' spacer with $R = 1.7$) and of hexanucleotides exhibiting frequencies higher than expected from their *R* ratios; and (ii) the exponential decrease in the average frequencies of the hexanucleotide classes with decreasing *R* ratio (Fig. 4), the slope in a semi-logarithmic plot depending upon the *R* ratio.

The simulation experiments indicate that the simplest model for the mechanism of formation of the AT spacers during the evolution of the mitochondrial genome of yeast involved random rounds of duplication events (accompanied by inversions and translocations) acting first on an initial oligodeoxynucleotide and subsequently on its 'expansion sequences'. This initial oligodeoxynucleotide was characterized by a particular *R* ratio. This sequence constraint, propagated in the expansion process, is analogous to the initial *R* ratio of the four A/T dinucleotides used to build the 'simulated' AT spacer. As in the latter case, the initial *R* ratio is different from the final one because the build-up of the sequence involves a drift in ratio which is associated with 'border effects', namely to the contribution of dinucleotide junctions formed by the juxtaposition of duplicated oligodeoxynucleotides.

(c) The evolutionary origin of the AT spacers of the mitochondrial genome of *Saccharomyces cerevisiae* and *Torulopsis glabrata*.

A biologically important question is that of the identification of the initial oligodeoxynucleotide (discussed above) assumed to be the starting sequence for the formation of the ancestral AT spacer. In our view, the best candidate is a sequence derived from an ancestral promoter-replicator sequence also used for starting the RNA primers of nascent DNA chains in a prokaryote-like, compact mitochondrial genome, only made of genes.

(1) A comparison of *ori* sequences from *S. cerevisiae* with putative *ori* sequences from *T. glabrata* supports such a suggestion.

(i) The initiation of bidirectional replication of the mitochondrial DNA from *S. cerevisiae* is started by RNA primers copied from the *r* and *r** sequences which flank GC cluster C of *ori* sequences (Baldacci et al., 1984). The *r* sequence comprises a nonanucleotide identical to those used for initiating transcription of mitochondrial genes. Both sequences comprise a single G : C bp; moreover, the *r* sequence exhibits an *R* ratio close to 2. We have suggested (de Zamaroczy and Bernardi, 1986a) that the *r**-C-*r* system, the most essential part of the *ori* sequence, corresponds to a primitive *ori* sequence in which cluster C was not in the form of the present three penta-C's, but a single penta-C, which might have originated from a gene (the best candidate being either an rRNA or a tRNA gene).

(ii) The *ori* sequence(s) of *T. glabrata* has(have) not yet been identified experimentally. We know, however, that this genome does not contain anything even approaching the complex structure of the *ori* sequences of *S. cerevisiae* (which comprises three GC clusters), since no GC cluster is present in *T. glabrata*. This genome contains a number of nonanucleotides, TATAAGTAA (Clark-Walker et al., 1985) almost identical to those of *S. cerevisiae*, (A/T)TATAAGTA (see for a list de Zamaroczy and Bernardi, 1985) and of *Kluyveromyces lactis*, ATATAAGTA (Osinga and Tabak, 1982), and likewise used for initiating transcription. It is most likely, in our view, that at least one such nonanucleotide (or a related sequence) is included in the origin(s) of replication of *T. glabrata*.

In connection with this suggestion, it is interesting

TABLE X

Putative signal sequences for the initiation of DNA replication of mitochondrial genome from *Torulopsis glabrata*

(a) *oli1*...ATACTAATTA . . 44-nt . . TATACTAATTATATATTTAATATATTATATA

ATACTAATTATATATTTAATATATTATATAATACTAATTA . . . *oxi1*

(b) *oxi1*...TATAATATTCTTATTTTATATATATAAGTAATAATA *phe*-tRNA

Line (a): the four decanucleotides (Clark-Walker et al., 1985) are underlined. Two of them are part of two 30-nt-long contiguous direct repeats (underlined with a broken line). Overlined sequences are homologous to the *r* stretches of *ori* sequences from *S. cerevisiae*. An additional T nucleotide present in *S. cerevisiae* is indicated.
Line (b): the processing dodecanucleotide of the *oxi1* gene and the promoter nonanucleotide of the *phe* tRNA are underlined. The sequences identical to the nonanucleotides and to the *r* stretches of *ori* sequences from *S. cerevisiae* are overlined. A single base difference present in *S. cerevisiae* is indicated. All sequence data are from Clark-Walker et al. (1985).

to note that one of these promoter sequences (that of the *phe* tRNA) is part of a 16-nt sequence which is identical, but for 1 nt, to sequence *r* (Table X). On the other hand, four repeated decanucleotides in an apparently non-transcribed region have been suggested to correspond to the *ori* sequence(s) of *T. glabrata* (Clark-Walker et al., 1985). Two of such decanucleotides are located in two contiguous, repeated 30-nt sequences, which are similar in their first part to the *r* sequence (Table X).

In conclusion, it is most likely in our opinion that the *ori* sequence(s) of *T. glabrata* correspond(s) to an A + T-rich sequence which comprises a nonanucleotide sequence and, possibly, an *r* sequence. Obviously, this putative *ori* sequence is similar, except for cluster C, to the primitive *ori* sequence of *S. cerevisiae* and not too far from the postulated ancestral replicator-promoter sequence. Another point of interest in the mitochondrial genome of *T. glabrata* is that the dodecanucleotide processing sites (which are identical to those of *S. cerevisiae*), are very close to the following nonanucleotide promoters, from which they could be derived.

At this point, it should be stressed that all the similarities between the mitochondrial genomes of *S. cerevisiae* and *T. glabrata* (nonanucleotide promoters, dodecanucleotide processing sites, AT spacers, and *var1* genes) definitely point to a closer phylogenetical relationship than expected from their current taxonomical positions.

(2) Some additional points are worth considering in connection with our proposed identification of the initial oligodeoxynucleotide.

(i) A sequence expansion around an origin of replication is most likely not to interfere with genome functions, as shown by the fact that variability in the sequence and size of the compact mitochondrial genomes of animals only exists around the replication origin (D-loop region).

(ii) The 1077-nt-long A + T-rich region harboring the DNA replication initiation sequence of the mitochondrial genome of *Drosophila yakuba* (Clary and Wolstenholme, 1985) does not contain any sequence closely resembling the nonanucleotide promoter of yeast. The *D. yakuba* sequence is characterized by an asymmetry of A and T, and of G and C, the latter two bases being more abundant than in the 'compiled' AT spacer of *S. cerevisiae*. An analysis of A/T di- and trinucleotides reveals an essentially statistical distribution, except for a 10% excess of AAA and TTT (not shown). This sequence might therefore be derived from another A + T-rich initial oligodeoxynucleotide, possibly also involved in the initiation of replication.

(iii) The suggestion that at least one of the *T. glabrata* nonanucleotides is used as a replication origin may also apply to the initiation of replication in *ori*° petites from *S. cerevisiae* (which carry mitochondrial genomes without any canonical *ori* sequence); indeed, we have observed a nonanucleotide

sequence on one or the other strand in four out of five completely sequenced *ori*$^{-}$ petite genomes (Goursot et al., 1982); the fifth contains a nonanucleotide only presenting a single base difference from the canonical ones.

(iv) The possibility that AT spacers have a nuclear origin was also explored, since nucleotide sequence exchanges between nucleus and mitochondria are known (Butow et al., 1985). The results from the 57 flanking sequences (excluding UTS when defined) of 41 nuclear genes of *S. cerevisiae* (as extracted from GenBank Release 44, August 1986) corresponding to a total of about 30000 nt showed, however, no common features in composition and A/T dinucleotide distribution with the AT spacers. Indeed, the 'compiled' nuclear sequences are characterized by a much higher GC level (37%) and exhibit an excess of AA and TT (14% and 19%, respectively) and a shortage of AT and TA (9% and 19%, respectively) compared to statistical expectation. Moreover, no excess of CC and GG was found. The only feature in common with the 'compiled' AT spacer is the equimolarity of A and T, and of G and C, respectively. Interestingly, the vast majority of individual flanking sequences exhibit the same deviations of the A/T dinucleotides as the overall sequences in spite of large differences in size and composition (not shown). This last point (to be discussed in more detail elsewhere) leads again to postulating a mechanism of sequence expansion analogous to that just described for the AT spacers of the mitochondrial genome of *S. cerevisiae*.

(d) The evolutionary origin of the *var1* gene

(1) The first comparison of long (1 kb) mitochondrial sequences from *S. cerevisiae* involved the *var1* and the *ori*1 'loci' (Bernardi and Bernardi, 1980). It revealed that many sequences of the two segments were identical, in agreement with the idea that spacer sequences are built according to the same pattern all over the mitochondrial genome. When the existence of the *var1* gene was substantiated (Hudspeth et al., 1982), this led to the proposal of an endogenous origin for the gene (Bernardi, 1983).

Detailed evidence in favor of a common endogenous origin is provided by the present work for the *var1* genes of both *S. cerevisiae* and *T. glabrata*. Indeed, in spite of differences (which will be accounted for in the following section) in base composition and in the frequency deviations of di- to hexanucleotides, the *var1* genes from both *S. cerevisiae* and *T. glabrata* exhibit the characteristic sequence pattern of the AT spacers (see RESULTS, section **d**, and Tables VII and VIII). Needless to say, the endogenous origin of *var1* accounts for the fact that this gene is not found in the mitochondrial genomes of other fungi (e.g., *Neurospora crassa*, *Aspergillus nidulans* and *Schizosaccharomyces pombe*), with the exception of *T. glabrata* (Clark-Walker et al., 1985), a yeast containing intergenic AT spacers clearly related to that of *S. cerevisiae*. This situation is in sharp contrast with the high interspecific conservation of the genes for respiratory proteins.

(2) A different proposal concerning the evolutionary origin of the *var1* gene of *S. cerevisiae* was that of a second, independent colonization event (Hudspeth et al., 1982; Butow et al., 1985). This hypothesis was put forward to explain the different codon usage of *var1* (as well as of the ORFs) compared to mitochondrial genes for respiratory proteins. Such a difference does not, however, require the explanation provided, in view of the well demonstrated possibility of change in codon usage during evolution and of its dependence upon compositional constraints, which may even be different in different regions of the same genome (Bernardi et al., 1985; Bernardi and Bernardi, 1986a,b).

In view of the absence of any other evidence for an exogenous origin of the *var1* gene, and of the presence of convincing evidence for an endogenous origin, the hypothesis of Hudspeth et al. (1982) should be abandoned.

(e) The mechanism of formation of the *var1* gene

(1) According to the recruitment hypothesis (Zassenhaus and Butow, 1983) 'the *var1* gene may have arisen by recombination between highly repeated spacer DNA sequences, bringing together pre-existing domains into a functional transcription unit, followed perhaps by transposition of GC clusters into the gene' (Ainley et al., 1985). These domains of 'short unexpressed open reading frames... may have existed earlier as individual genes or coding units' (Butow et al., 1985). Two main arguments were given in support of this hypothesis.

An argument was that sequences in and around *b1* and *b2* (the two runs of AAT repeats, whose numbers differ in different strains), are highly repeated in the AT spacer, whereas other longer (> 15 nt) sequences representing the majority of the gene are unique or only occasionally repeated elsewhere in the genome (Butow et al., 1985). The authors considered this as a special feature of the *var1* gene, which supported a 'recombinational shuffling' model along the line of 'exon shuffling' (Gilbert, 1978; Darnell, 1978). We have shown, however, that this situation is not unique to the *var1* gene, but is also regularly found in the AT spacers. Indeed, runs of short repeats, varying in number in different strains, exist in the AT spacers (de Zamaroczy and Bernardi, 1986b), and the majority of A/T 14-nt sequences are only present in one to four copies (see RESULTS, section **b**(4)); the most frequent ones in fact comprise the alternating AT sequences and the AAT repeat runs.

Another argument was that 'the location of GC clusters within the mitochondrial genome is highly variable among different *S. cerevisiae* strains' (Ainley et al., 1985). This argument was based on an inter-strain comparison of two GC clusters in and before the *var1* gene (Hudspeth et al., 1984) and of two other previously described GC clusters in the *15S-RNA* and *21S-RNA* genes, respectively, and was taken as an additional evidence of a high level of the site-specific recombination events which were considered to be responsible for the formation of the *var1* gene. In fact, only occasional differences (presence/absence or partial rearrangement) of 'allelic' GC clusters have been found in pairwise comparisons of many strains (de Zamaroczy and Bernardi, 1986a,b). Such differences appear to be rare events and not a general rule for the 220 GC clusters of the 'long' mitochondrial genome of yeast. As we have previously shown both the total number of GC clusters and their physical distribution in the genome are very close in different strains (Prunell et al., 1977). In other words, there is no evidence for the large-scale genome rearrangements required by the 'recombinational shuffling' hypothesis. Other objections to the hypothesis of recruitment of previously coding sequences are that the short ORFs regularly present along the genome (see section... below) have a sequence pattern of the AT spacer type (except for the major intergenic ORFs, which show, however, a sequence pattern different from *var1*; see section **f** below). In other words, no 'earlier' coding units in the AT spacers appear to have the special *var1* features.

(2) The linkage-phasing hypothesis we propose here for the mechanism of formation of the *var1* gene is (a) that this gene initially corresponded to an AT spacer segment located in a transcribed region (such situation is the rule in the mitochondrial genome of yeast); (b) that short ORFs in this region became linked and put in phase as the result of point mutations; (c) that, at a certain point in time, this transcript could be translated into a protein that could play a role in the small sub-unit of mitochondrial ribosomes; and (d) that the gene underwent an elongation as the result of point mutations and of GC cluster insertion which extended the ORF.

This proposal is supported by the fact that short ORFs potentially coding for 30–50 aa (mainly beginning with AUA as initiation codons) are common in intergenic sequences. Linkage and phasing of several such ORFs can be achieved with 1–2 nt deletions/additions and point mutations in an approximate ratio of 2 : 1 (namely the ratio detected in polymorphic sequences; de Zamaroczy and Bernardi, 1986b). For instance, an ORF potentially coding for 591 aa can be 'constructed' in the 1.7-kb region comprised between *ori*3 and *ori*4 with only 1.8% base changes, namely with the same divergence level found in interstrain comparisons of intergenic sequences. Because of the high degree of homology of intergenic sequences in the size range of 2–5 kb, similar constructions are possible elsewhere.

(3) The sequence features of the *var1* genes of *S. cerevisiae* and *T. glabrata* are due to specific changes acting on the AT spacer sequence which originated the *var1* gene.

The stronger similarity, at the sequence pattern level, between the two *var1* genes compared to the two intergenic sequences indicates that the appearance of the special common features of the genes has preceded the separation of the two species. These special features comprise an asymmetric distribution of A and T (A being in excess over T by 10% on the non-coding strand). Alternating AT sequences have therefore been enriched in A by T → A transversions, and/or by A additions, and/or by T deletions; this is illustrated (Table VIII) by the frequency excess for AATAAT, ATAATA and TAATAA, due to the shortage of hexanucleotides only made of, or enrich-

ed in, alternating AT. (The excess due to the contribution of the two AAT inserts in the *var1* gene of *S. cerevisiae* is neglected here.) In the case of the *var1* gene, however, the A enrichment explains that one of the two sets of complementary sequences is systematically less frequent. It should be noted that the same mechanisms are also responsible for the polymorphism of intergenic sequences, but in the AT spacer A↔T transversions are bidirectional and therefore lead to no change in sequence pattern.

Interestingly, the 15% divergence (Ainley et al., 1985) between the two *var1* genes of *S. cerevisiae* and *T. glabrata* (mainly due to bidirectional A↔T transversions, which are three times as frequent as transitions) is not accompanied by an additional change in sequence pattern nor in the AT levels of individual codon positions; it is accompanied, however, by many amino acid changes (leading to increased levels of methionine, asparagine and aspartic acid in the *var1* protein of *S. cerevisiae*; Ainley et al., 1985) which stress the lack of strong functional constraints in the *var1* protein.

(f) The other intergenic sequences of *S. cerevisiae*

(1) The *l* stretches of *ori* sequences are clearly related to the AT spacers from which they are derived. Like the *var1* gene, these sequences exhibit some special features which are associated, in all likelihood, with their function (this has been discussed elsewhere; see de Zamaroczy et al., 1984).

(2) The UTS sequences comprise both AT spacers and GC clusters; the former are so closely similar in sequence pattern to the intergenic AT spacers that they cannot be distinguished from them.

(3) The intergenic ORFs belong to two different sets. ORFs 1, 2 and 4 are in part distinct in sequence properties from the AT spacers, whereas they are similar to some intronic ORFs in both composition and di- and trinucleotide distribution; these sequences also present a limited, patchy homology at the amino acid level with some intronic ORFs (Michel, 1984; Séraphin et al., 1985). The evolutionary origin of these major ORFs will be discussed elsewhere, in connection with that of intronic ORFs (M. de Z. and G.B.; paper in preparation). In contrast, ORF3 and ORF5 are very closely related to the AT spacer sequences. If ORF5 corresponds to a mitochondrial gene, as it is possible (Colin et al., 1985), its origin would then be the same as that of *var1*, its closer similarity to the AT spacer suggesting a more recent origin.

(g) The biological functions of AT spacers

(1) Mitochondrial recombination, so far, has been shown to involve AT spacers (and GC clusters) and is an extremely frequent event which follows crossing of wild-type *S. cerevisiae* cells (Fonty et al., 1978). Such recombination events do not lead to any long-range rearrangements of the mitochondrial genome, as far as one can judge from detailed restriction analysis, but only to short deletion/additions at crossing-over sites, which generate polymorphisms. Similar recombination events are responsible for the generation of the defective 'petite' genomes, in which case excision sequences located in both the AT spacers and the GC clusters have been defined (de Zamaroczy et al., 1983).

The role played in recombination by AT spacers (and GC clusters), is important in evolution and certainly underlies the gene order changes observed in a number of mitochondrial genomes and in particular in those of *S. cerevisiae* and *T. glabrata* (Clark-Walker, 1985). Another evolutionary role played by AT spacers consists in the generation of genes, like *var1*, as well as of regulatory sequences like the *l* stretches of *ori* sequences and UTS.

(2) The major physiological role played by AT spacers (and GC clusters) has to do, in all likelihood, with the structure of the mitochondrial genome and in a direct or an indirect way, with the function of its regulatory elements.

As far as the structural role is concerned, it is worthwhile stressing that in *S. cerevisiae* the GC level of intergenic sequences is practically constant down to DNA segment sizes as low as 3000 nt. The higher GC level of intergenic sequences, 14%, compared to that of AT spacers, 5%, is due to the contribution of GC clusters, which leads, in spite of local variations in number and classes, to such constant value. The absence of GC clusters in the *T. glabrata* genome appears to correlate with the shorter size of AT spacers compared to those of *S. cerevisiae*.

As for the functional role of AT spacers, the suggestion is that regulatory elements may be modulated by the genome configuration. The general idea is that signal sequences used in initiating transcription,

replication, splicing, processing, recombination, are necessary but not sufficient for the different functions involved, and that the neighboring sequences and/or the secondary/tertiary structure of the region, and/or its 'chromatin' structure also play a role. This view is supported by both general and specific arguments.

Among the general arguments, the very strict conservation in amount of intergenic sequences of mitochondrial genomes from different wild-type strains, in spite of the frequent excision events occurring in them, strongly support a 'functional' role for those sequences (Bernardi, 1983). The interstrain conservation of the primary structure of intergenic sequences (de Zamaroczy and Bernardi, 1985; 1986a,b) provides an additional argument along the same line.

Among the specific arguments, some examples concerning nonanucleotides which are not used as transcription starts in spite of being 'canonical' in sequence have been quoted elsewhere (de Zamaroczy and Bernardi, 1986a). Likewise, seven promoter sequences and one dodecamer processing sequence (Osinga et al., 1984b) exist on the non-transcribed strand (which is barely used as a template), but only one of them, the promoter of the *thr1* gene, is actually utilized by the mitochondrial RNA polymerase. On the other hand, promoter sequences which are apparently silent in the genome of wild-type cells can become active in the defective genome of some 'petite' mutants; such is the case of the *cys* tRNA gene promoter (Frontali et al., 1985). Moreover, AT spacers transcribed into 5′- and 3′-untranslated sequences of mature RNAs play a role in the processing (Zassenhaus et al., 1984; see Dieckmann et al., 1984; Müller et al., 1984) and possibly in the stability and splicing of these RNAs. For instance, mutations in the flanking regions of the *var1* gene lead to a lack of expression of this gene (Butow et al., 1985).

The availability of mitochondrial genomes like that of *T. glabrata* which contain AT spacers clearly evolutionarily related to those of *S. cerevisiae*, but differing in amount as well as in some features, like the absence of GC clusters, may be useful models to compare regulatory properties of the two genomes. Such a comparison could also show that the expansion-contraction process affecting the intergenic sequences is correlated with gains/losses in regulatory properties.

ACKNOWLEDGEMENTS

We are particularly grateful to Claude Mugnier for his constant help with computer programs and especially the simulation programs used in the study of AT spacers. Sequence data treatments were performed using computer facilities at CITI2 in Paris, with the help of the French Ministère de la Recherche et de la Technologie.

We thank Martine Brient for typing this manuscript, and Philippe Breton for the artwork.

REFERENCES

Ainley, W.M., Macreadie, I.G. and Butow, R.A.: *var1* gene on the mitochondrial genome of *Torulopsis glabrata*. J. Mol. Biol. 184 (1985) 565–576.

Baldacci, G., Chérif-Zahar, B. and Bernardi, G.: The initiation of DNA replication in the mitochondrial genome of yeast. EMBO J. 3 (1984) 2115–2120.

Bernardi, G., Carnevali, F., Nicolaieff, A., Piperno, G. and Tecce, G.: Separation and characterization of a satellite DNA from a yeast cytoplasmic 'petite' mutant. J. Mol. Biol. 37 (1968) 493–505.

Bernardi, G., Faurès, M., Piperno, G. and Slonimski, P.P.: Mitochondrial DNA's from respiratory-sufficient and cytoplasmic respiratory-deficient mutant yeast. J. Mol. Biol. 48 (1970) 23–42.

Bernardi, G. and Timasheff, S.N.: Optical rotatory dispersion and circular dichroism. Properties of yeast mitochondrial DNA's. J. Mol. Biol. 48 (1970) 43–52.

Bernardi, G. and Bernardi, G.: Repeated sequences in the mitochondrial genome of yeast. FEBS Lett. 115 (1980) 159–162.

Bernardi, G.: Evolutionary origin and the biological function of non-coding sequences in the mitochondrial genome of yeast. In Slonimski, P.P., Borst, P. and Attardi, G. (Eds.), Mitochondrial Genes. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982, pp. 269–278.

Bernardi, G.: Genome instability and the selfish DNA issue. Folia Biol. 29 (1983) 82–92.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F.: The mosaic genome of warm-blooded vertebrates. Science 228 (1985) 953–958.

Bernardi, G. and Bernardi, G.: The human genome and its evolutionary context. Cold Spring Harbor Symp. Quant. Biol. 51 (1986a) in press.

Bernardi, G. and Bernardi, G.: Compositional constraints and genome evolution. J. Mol. Evol. 24 (1986b) 1–11.

Blanc, H. and Dujon, B.: Replicator regions of the yeast mitochondrial DNA responsible for suppressiveness. Proc. Natl. Acad. Sci. USA 77 (1980) 3942–3946.

Bonitz, S.G., Homison, G., Thalenfeld, B.E., Tzagoloff, A. and Nobrega, F.G.: Assembly of the mitochondrial membrane system. Processing of the apocytochrome *b* precursor RNAs

in *Saccharomyces cerevisiae* D273-10B. J. Biol. Chem. 257 (1982) 6268–6274.

Butow, R.A., Perlman, P.S., Grossman, L.I.: The unusual *var1* gene of yeast mitochondrial DNA. Science 228 (1985) 1496–1501.

Christianson, T. and Rabinowitz, M.: Identification of multiple transcriptional initiation sites on the yeast mitochondrial genome by in vitro capping with guanylyltransferase. J. Biol. Chem. 258 (1983) 14025–14033.

Clark-Walker, G.D.: Basis of diversity in mitochondrial DNAs. In Cavalier-Smith, T. (Ed.), The Evolution of Genome Size, Wiley, New York, Chapter 10, 1985, pp. 277–297.

Clark-Walker, G.D., McArthur, C.R. and Sriprakash, K.S.: Location of transcriptional control signals and transfer RNA sequences in *Torulopsis glabrata* mitochondrial DNA. EMBO J. 4 (1985) 465–473.

Clary, D.O. and Wolstenholme, D.R.: The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. J. Mol. Evol. 22 (1985) 252–271.

Colin, Y., Baldacci, G. and Bernardi, G.: A new putative gene in the mitochondrial genome of *Saccharomyces cerevisiae*. Gene 36 (1985) 1–13.

Corneo, G., Moore, C., Sanadi D.R., Grossman L.I. and Marmur, J.: Mitochondrial DNA in yeast and some mammalian species. Science 151 (1966) 687–689.

Cosson, J. and Tzagoloff, A.: Sequence homologies of (guanosine + cytidine)-rich regions of mitochondrial DNA of *Saccharomyces cerevisiae*. J. Biol. Chem. 254 (1979) 42–43.

Darnell, J.E.: Implications of RNA-RNA splicing in evolution of eukaryotic cells. Science 202 (1978) 1257–1260.

de Zamaroczy, M. and Bernardi, G.: Sequence organization of the mitochondrial genome of yeast — a review. Gene 37 (1985) 1–17.

de Zamaroczy, M. and Bernardi, G.: The GC clusters of the mitochondrial genome of yeast and their evolutionary origin. Gene 41 (1986a) 1–22.

de Zamaroczy, M. and Bernardi, G.: The primary structure of the mitochondrial genome of *Saccharomyces cerevisiae* — a review. Gene 47 (1986b) 155–177.

de Zamaroczy, M., Baldacci, G. and Bernardi, G.: Putative origins of replication in the mitochondrial genome of yeast. FEBS Lett. 108 (1979) 429–432.

de Zamaroczy, M., Marotta, R., Faugeron-Fonty, G., Goursot, R., Mangin, M., Baldacci, G. and Bernardi, G.: The origins of replication of the yeast mitochondrial genome and the phenomenon of suppressivity. Nature 292 (1981) 75–78.

de Zamaroczy, M., Faugeron-Fonty, G. and Bernardi, G.: Excision sequences in the mitochondrial genome of yeast. Gene 21 (1983) 193–202.

de Zamaroczy, M., Faugeron-Fonty, G., Baldacci, G., Goursot, R. and Bernardi, G.: The *ori* sequences of the mitochondrial genome of a wild-type yeast strain: number, location, orientation and structure. Gene 32 (1984) 439–457.

Dieckmann, C.L., Koerner, T.J. and Tzagoloff, A.: Assembly of the mitochondrial membrane system. *CBP1*, a yeast nuclear gene involved in 5′ end processing of cytochrome *b* PRE-mRNA. J. Biol. Chem. 259 (1984) 4722–4731.

Edwards, J.C., Osinga, K.A., Christianson, T., Hensgens, L.A.M., Janssens, P.M., Rabinowitz, M. and Tabak, H.F.: Initiation of transcription of the yeast mitochondrial gene coding for ATPase subunit 9. Nucl. Acids Res. 11 (1983) 8269–8282.

Ehrlich, S.D., Thiery, J.-P. and Bernardi, G.: The mitochondrial genome of wild-type yeast cells, III. The pyrimidine tracts of mitochondrial DNA. J. Mol. Biol. 65 (1972) 207–212.

Fonty, G., Goursot, R., Wilkie, D. and Bernardi, G.: The mitochondrial genome of wild-type yeast cells, VII. Recombination in crosses. J. Mol. Biol. 119 (1978) 213–235.

Frontali, L., Francisci, S., Palleschi, C., Stifani, S. and Zennaro, E.: Transcription initiation and processing sites in the tRNA region of the yeast mitochondrial genome. In Quagliariello, E., Slater, E.C., Palmieri, F., Saccone, C., and Kroon, A.M. (Eds.), Achievements and Perspectives of Mitochondrial Research, Vol. II: Biogenesis. Elsevier, Amsterdam, 1985, pp. 203–214.

Gaillard, C. and Bernardi, G.: The nucleotide sequence of the mitochondrial genome of a spontaneous 'petite' mutant of yeast. Mol. Gen. Genet. 174 (1979) 335–337.

Gaillard, C., Strauss, F. and Bernardi, G.: Excision sequences in the mitochondrial genome of yeast. Nature 283 (1980) 218–220.

Gaillard, C., Doly, J., Cortadas, J. and Bernardi, G.: The primary structure of bovine satellite 1.715. Nucl. Acids Res. 9 (1981) 6069–6082.

Gilbert, W.: Why genes in pieces? Nature 271 (1978) 501.

Goursot, R., de Zamaroczy, M., Baldacci, G. and Bernardi, G.: Supersuppressive 'petite' mutants of yeast. Curr. Genet. 1 (1980) 173–176.

Goursot, R., Mangin, M. and Bernardi, G.: Surrogate origins of replication in the mitochondrial genomes of *ori*° petite mutants of yeast. EMBO J. 1 (1982) 705–711.

Hudspeth, M.E.S., Ainley, W.M., Shumard, D.S., Butow, R.A. and Grossman, L.I.: Location and structure of the *var1* gene on yeast mitochondrial DNA: nucleotide sequence of the *40.0* allele. Cell 30 (1982) 617–626.

Hudspeth, M.E.S., Vincent, R.D., Perlman, P.S., Shumard, D.S., Treisman, L.O. and Grossman, L.I.: Expandable *var1* gene of yeast mitochondrial DNA: in-frame insertions can explain the strain-specific protein size polymorphisms. Proc. Natl. Acad. Sci. USA 81 (1984) 3148–3152.

Michel, F.: A maturase-like coding sequence downstream of the *oxi2* gene of yeast mitochondrial DNA is interrupted two GC clusters and a putative end-of-messenger signal. Curr. Genet. 8 (1984) 307–317.

Mounolou, J.-C., Jakob, H. and Slonimski, P.P.: Mitochondrial DNA from yeast 'petite' mutants: specific changes of buoyant density corresponding to different cytoplasmic mutations. Biochem. Biophys. Res. Commun. 24 (1966) 218–224.

Müller, P.P., Reif, M.K., Zonghou, S., Sengstag, C., Mason, T.L. and Fox, T.D.: A nuclear mutation that post-transcriptionally blocks accumulation of a yeast mitochondrial gene product can be suppressed by a mitochondrial gene rearrangement. J. Mol. Biol. 175 (1984) 431–452.

Osinga, K.A. and Tabak, H.F.: Initiation of transcription of genes for mitochondrial ribosomal RNA in yeast: comparison

of the nucleotide sequence around the 5′-ends of both genes reveals a homologous stretch of 17 nucleotides. Nucl. Acids Res. 10 (1982) 3617–3626.

Osinga, K.A., De Vries, E., Van der Horst, G.T.J. and Tabak, H.F.: Initiation of transcription in yeast mitochondria analysis of origins of replication and of genes coding for a messenger RNA and a transfer RNA. Nucl. Acids Res. 12 (1984a) 1889–1900.

Osinga, K.A., De Vries, E., Van der Horst, G. and Tabak, H.F.: Processing of yeast mitochondrial messenger RNAs at a conserved dodecamer sequence. EMBO J. 3 (1984b) 829–834.

Piperno, G., Fonty, G. and Bernardi, G.: The mitochondrial genome of wild-type yeast cells. II. Investigations on the compositional heterogeneity of mitochondrial DNA. J. Mol. Biol. 65 (1972) 191–205.

Prunell, A. and Bernardi, G.: The mitochondrial genome of wild-type yeast cells. IV. Genes and spacers. J. Mol. Biol. 86 (1974) 825–841.

Prunell, A. and Bernardi, G.: The mitochondrial genome of wild-type yeast cells, VI. Genome organization. J. Mol. Biol. 110 (1977) 53–74.

Prunell, A., Kopecka, H., Strauss, F. and Bernardi, G.: The mitochondrial genome of wild-type yeast cells. V. Genome evolution. J. Mol. Biol. 110 (1977) 17–52.

Schildkraut, C.L., Marmur, J. and Doty, P.: Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. J. Mol. Biol. 4 (1962) 430–443.

Séraphin, B., Simon, M. and Faye, G.: A mitochondrial reading frame which may code for a maturase-like protein in *S. cerevisiae*. Nucl. Acids Res. 13 (1985) 3005–3014.

Szybalski, W.: Use of cesium sulfate for equilibrium density gradient centrifugation. Methods in Enzymology 12B (1968) 330–360.

Tewari, K.K., Jayaraman, J. and Mahler, H.R.: Separation and characterization of mitochondrial DNA from yeast. Biochem. Biophys. Res. Comm. 21 (1965) 141–148.

Thalenfeld, B.E., Hill, J. and Tzagoloff, A.: Assembly of the mitochondrial membrane system. Characterization of the *oxi2* transcript and localization of its promoter in *Saccharomyces cerevisiae* D273-10B. J. Biol. Chem. 258 (1983) 610–615.

Van Kreijl, C.F. and Bos, J.L.: The repeating nucleotide sequence in the repetitive mitochondrial DNA from a 'low-density' petite mutant of yeast. Nucl. Acids Res. 4 (1977) 2369–2388.

Zassenhaus, H.P. and Butow, R.A.: Functions of nongenic DNA in yeast mitochondria. In Schweyen, R.J., Wolf, K. and Kaudewitz, F. (Eds.), Mitochondria 1983: Nucleo-Mitochondrial Interactions, De Gruyter, Berlin, 1983, pp. 95–106.

Communicated by M.R. Culbertson.