Gene, 122 (1992) 91–99

GENE 06836

The mosaic organization of the mitochondrial introns of *Saccharomyces cerevisiae*: features and evolutionary origins

(Oligodeoxyribonucleotide analysis; intronic closed reading frames; intronic open reading frames; intergenic sequences)

Miklos de Zamaroczy* and Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Received by C. Saccone: 28 August 1992; Accepted: 2 September 1992; Received at publishers: 10 September 1992

SUMMARY

The introns of three genes (*oxi3*, *cob* and *21S*) from the mitochondrial (mt) genome of *Saccharomyces cerevisiae* contain closed reading frames (CRFs). In the present work, we have analyzed these sequences in their oligodeoxyribonucleotide (oligo; isostich) patterns. We have shown that the relative amounts of di- to hexanucleotides, when compared to random sequences having the same sizes and compositions, exhibit the same deviations as the intergenic noncoding sequences of the mt genome (except for the CRFs from *21S* intron). In contrast, intronic open reading frames (ORFs) showed oligo patterns which were generally quite distinct from those of CRFs, although some similarities could be detected in some cases (especially for aI5α). The mt introns of yeast, therefore, are endowed with a mosaic structure, in which CRFs derive from mt intergenic sequences, whereas ORFs have a different origin (indicated as exogenous by other evidences) yet show, in some cases, the effects of 'sequence assimilation' with CRFs.

INTRODUCTION

Intergenic sequences represent 63% of the mt 'long' (85 kb) genome of *S. 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 coding sequences. 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 these spacers are characterized by an extremely high level of short-sequence repetitiveness and by a characteristic sequence pattern; the frequencies of A/T oligos having the same size (isostichs) conspicuously deviate from statistical expectations, and exponentially decrease when their (AT+TA)/(AA+TT) ratio, R , decreases (de Zamaroczy and Bernardi, 1987).

The sequence features of the AT spacers indicate (de Zamaroczy and Bernardi, 1986a) that they were built in evolution by an expansion process which involved an initial oligo apparently corresponding to, or arisen from, an ancestral promoter-replicator sequence. An origin from a primitive *ori* sequence was also postulated for the 196 GC

Correspondence to: Dr. G. Bernardi, Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France. Tel. (33-1) 43 29 58 24; Fax (33-1) 44 27 79 77; e-mail Bernardi@FRCIT151.BITNET.

* Permanent address: Unité de Physiologie Cellulaire (CNRS URA 1300), Département des Biotechnologies, Institut Pasteur, 25, Rue du Dr. Roux, 75724 Paris Cedex 15, France. Tel.(33-1) 45 68 80 00, ext. 7655; Fax (33-1) 45 68 87 90.

Abbreviations: bp, base pair(s); CRF, closed reading frame(s); CSE, conserved sequence element(s); GC,% of guanine+cytosine; IGS, internal guide sequence; kb, kilobase(s) or 1000 bp; mt, mitochondrial; nt, nucleotide(s); oligo, oligodeoxyribonucleotide; ORF, open reading frames; *ori*, origin(s) of DNA replication; *S.*, *Saccharomyces*; TSS, typical secondary structure.

clusters present in the 90% of the mt genome which have already been sequenced. The vast majority of GC clusters is located in intergenic sequences (including *ori* sequences, intergenic ORFs and the gene *var1*, which arose itself from an intergenic spacer sequence) and in intronic CRFs.

Three genes, *oxi3*, *cob* and *21S*(*LSU-rRNA*) from the 'long' (85 kb) mt genome of *S. cerevisiae* comprise nine, five and one intron(s), respectively. Apart from a stretch of about 250 nt, that is located in intron bI2 (and that corresponds to gap 13 in the available sequence of the mt genome; de Zamaroczy and Bernardi, 1986b), these introns are completely known in their primary structure, and comprise long ORFs, except for introns aI6, bI1 and bI5, which only consist of CRFs, and for the very short (20–30 bp) introns aI7 and aI8. These ORFs are contiguous to, and in reading frame phase with, those of preceding exons, except for ORFs from introns rI and aI5β. All introns from the mt genome of *S. cerevisiae* comprise CRFs, corresponding to a total of 5300 bp, namely 6.3% of the 'long' genome.

Two groups, I and II, of mt introns from *S. cerevisiae* and other fungi have been defined on the basis of TSS of the corresponding transcripts (Michel et al., 1982; Michel and Dujon, 1983; Davies et al., 1982; Waring and Davies, 1984; see for recent reviews Michel et al., 1989; Cech, 1988; 1990; Burke, 1988; Dujon, 1989). All group-I introns contain the short CSE: P, Q, R, S, E and E′, and an IGS has been proposed to bring the two splice sites together. The group I mt introns share conserved sequences as well as structural elements with the nuclear rRNA gene intron from *Tetrahymena thermophila*, which undergoes a self-splicing reaction (Cech et al., 1981; 1983; Michel and Dujon, 1983). A common set of paired regions (P1-P9) form the *core* of the structure. Extra sequences, usually including long ORFs, form stem-and-loop structures that are peripheral to the core and are much more variable between introns. Group-II introns show a distinct secondary structure, which includes a conserved 38-nt stem-and-loop element near their 3′ ends. They are unusual in that the excised introns form stable lariat structures (Van der Veen et al., 1986). Altogether, in the mt genome of *S. cerevisiae* , there are only few members of this group, whereas the majority of introns belong to group I.

In the present work, we have investigated the sequences of intronic CRFs and ORFs and shown that their features indicate that they derived from intergenic sequences in the first case, and from other (exogenous) sequences in the second. Sequences were analyzed in their oligo patterns by assessing the frequencies of several series of 2–6-nt long isostichs and comparing them with statistical expectations (see de Zamaroczy and Bernardi, 1987, for details on the approach used). The general principle of oligo comparisons in genome analysis above the nearest neighbor analysis level was introduced some 20 years ago (see Bernardi et al., 1973a,b) and has been recently applied by a number of laboratories (see for instance, Pietrokovski et al., 1990; Pizzi et al., 1991; Pesole et al., 1992).

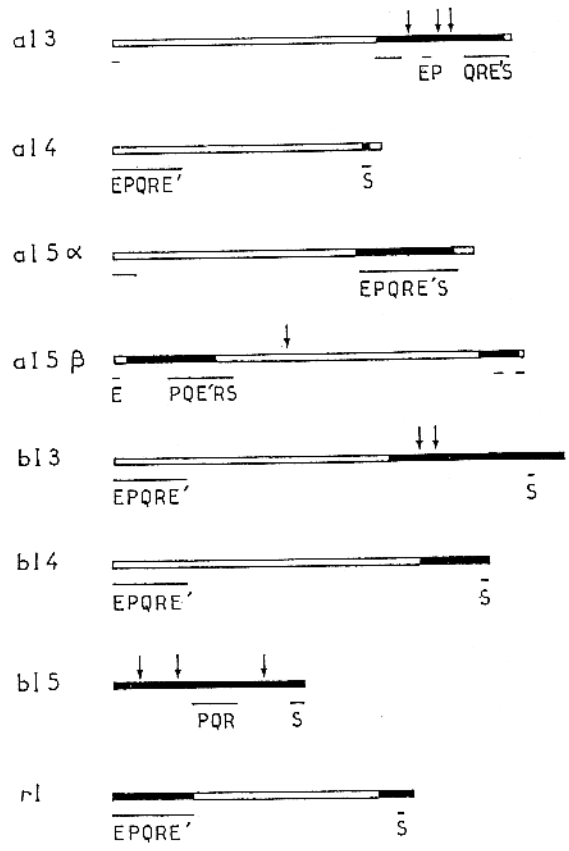

Fig. 1. Sequence organization of group-I introns from the mt genome of *S. cerevisiae*. Open bars, ORFs; blackened bars, CRFs. Thin lines, TSS. Capital letters indicate the location of conserved primary structure motifs (CSE). Arrows indicate the location of GC clusters, except for group *n*-type clusters.

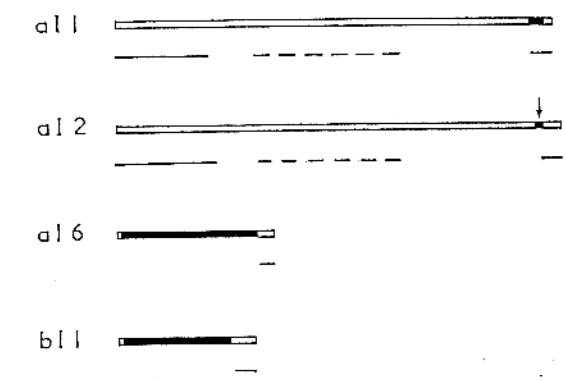

Fig. 2. Sequence organization of group II introns from the mt genome of *S. cerevisiae*. TSS are located within the thin lines (left end). Dashed lines indicate sequences with patched amino acid homology to retroviral sequences (Michel and Lang, 1985); the 3′ terminal (right) end sequences are conserved in primary structure. The arrow indicates the location of a GC cluster.

RESULTS AND DISCUSSION

(a) General sequence organization of mt introns

We have previously reported on the different intron patterns exhibited by different *S. cerevisiae* strains (de Zamaroczy and Bernardi, 1985), as well as on the sizes and the GC levels of introns. The corresponding primary structures (as well as all other available sequences) have been collated and presented by de Zamaroczy and Bernardi (1986b; see legend to Fig. 3).

Figs. 1 and 2 present the sequence organizations of mt introns of groups I and II, respectively, from *S. cerevisiae*, with their ORFs and CRFs; the location of sequences involved in the corresponding TSS are indicated. CSE have been localized (except for IGS) in the case of group-I introns. All GC clusters related with those of typical families from intergenic sequences (de Zamaroczy and Bernardi, 1986a) have also been localized (see section b).

Intronic CRFs can be classified in three different categories: (*i*) the very short CRFs (50–150 bp) from introns aI1, aI2 (group II; see Fig. 2) and aI4 (group I; see Fig. 1) are involved in the core-structure of TSS; (*ii*) the long CRFs, with sequences regularly involved in TSS of group I introns, comprise those present in introns aI3, aI5α, aI5β, bI5 and rI; (*iii*) the long CRFs, largely not involved in TSS, comprise those present in bI3, bI4 (group I) and aI6, bI1 (group II).

(b) Intronic GC clusters

Twenty-one GC clusters have been identified, classified and localized in introns (de Zamaroczy and Bernardi, 1986a); eleven of them are in the *oxi3* introns (intron aI6, formed by an 800-nt-long CRF, is exceptional in not containing any GC cluster, but only one G/C tetranucleotide), eight in the *cob* introns (bI), and two in the *21S* intron.

These GC clusters can be separated into two groups: (*i*) ten GC clusters, clearly related to families *a1–a4*, *c* and *v* (previously identified in intergenic sequences), are located in intronic CRFs, except for one *a1*-type cluster present in the ORF of intron aI5β. These GC clusters are present in either the 'major ORF loop', or in the variable loops of TSS, namely in sequences outside the core structure; (*ii*) eleven short atypical GC clusters belong to the *n* group; nine out of these eleven clusters (also located in CRFs) are involved in the core structure of TSS. These sequences do not comprise in most cases G and/or C stretches longer than tetranucleotides, whose presence is in agreement with the high GC levels of the harboring sequences.

TABLE I

GC levels, compositions and sizes (nt) of CRFs from mt introns of *Saccharomyces cerevisiae*[a]

	Group-I introns	G+C%	A	T	G	C	Size (nt)
oxi3	aI3	23.4	41.8	34.7	12.4	11.0	426
	aI5α	16.9	45.0	38.1	9.7	7.2	444
	aI5β1	12.3	48.7	39.0	7.2	5.1	374
	aI5β2	12.6	47.0	40.4	7.8	4.8	166
cob	bI2[b]	14.2	49.1	36.7	7.7	6.5	169
	bI3	9.8	43.7	46.4	4.7	5.1	487
	bI4	15.2	41.8	43.0	8.6	6.6	256
	bI5	11.2	48.1	40.7	5.9	5.3	646
21S	rI1	24.9	44.1	31.0	12.8	12.1	297
	rI2	12.3	40.6	47.1	8.0	4.3	138
	Group-II introns						
oxi3	aI1		n.d.				115
	aI2		n.d.				116
	aI6	17.7	43.8	38.6	10.4	7.3	811
cob	bI1	14.7	46.2	38.2	9.5	6.2	650
	Total[c]	15.3	45.2	39.5	8.6	6.7	4263

[a] Short intronic CRF sequences from aI7 and aI8 (≤30 bp long) and from aI4 (<60 bp long) have been neglected here and in the following tables. Base compositions are given as mol%. GC clusters of families *a1–a4*, *c* and *v* have not been taken into account, whereas those of group *n* involved in TSS have been.

[b] Gap 13 (about 250 bp) is present in bI2.

[c] Total does not comprise aI5β2 and the two CRFs of rI (rI1 and rI2; see section c). Some short putative ORFs, 15–115 nt long, in frame with the preceding or following exons, have been initially described for *oxi3* (Bonitz et al., 1980) and *cob* (Nobrega and Tzagoloff, 1980); since no indication has been provided so far that such sequences are functional, they are considered here together with CRFs.

TABLE II

Frequency differences of (*A*) dinucleotides and (*B*) A/T trinucleotides of CRFs from mt introns of *Saccharomyces cerevisiae*[a]

A	Introns	AA	TT	AT	TA	CC	GG	AC	CA
	aI3	− 11.4	− 25.8	+ 28.3	+ 29.7	+ 192	+ 40	− 23.9	− 43.5
	aI5α	− 8.9	− 6.9	+ 11.6	+ 12.8	+ 120	+ 22	− 15.6	− 28.1
	aI5β1	− 8.4	− 27.6	+ 18.4	+ 24.2	+ 167	+ 120	− 24.0	− 68.0
	aI5β2	+ 6.8	+ 0.6	− 7.4	− 4.2	—	—	− 21.7	− 21.7
	bI2	− 21.2	− 25.2	+ 28.9	+ 22.2	+ 200	0	+ 12.5	+ 50
	bI3	− 17.3	− 14.0	+ 18.7	+ 16.8	+ 100	+ 200	− 27.3	− 13.6
	bI4	− 10.3	− 10.8	+ 17.8	+ 4.4	+ 200	− 43	− 57.1	− 14.3
	bI5	− 11.3	− 15.7	+ 16.3	+ 14.8	+ 100	− 25	− 12.0	− 20.0
	rI1	− 3.1	− 5.2	+ 25.6	+ 5.8	+ 127	+ 25	− 35.9	− 11.3
	rI2	− 15.8	− 8.1	+ 14.7	+ 3.1	—	—	+ 29.4	+ 29.4
	aI6	− 10.4	− 16.1	+ 19.5	+ 16.0	+ 140	+ 46	− 15.6	− 28.1
	bI1	− 16.9	− 29.5	+ 26.0	+ 24.3	+ 125	+ 33	− 24.1	− 10.3
Total[b]		− 12.4	− 18.0	+ 20.4	+ 18.7	+ 167	+ 49	− 20.8	− 24.1
		AG	GA	TC	CT	TG	GT	CG	GC
Total[b]		− 10.0	+ 0.3	− 9.3	− 13.1	− 8.7	− 14.6	+ 56	+ 22
B		AAA	TTT	ATA	TAT	ATT	TTA	AAT	TAA
Total[b]		− 23.9	− 31.7	+ 40.7	+ 53.5	+ 4.2	+ 4.2	− 1.4	+ 6.2

[a] Frequency of differences are given as (found-expected)/expected, where 'found' refers to the frequencies in the CRF and 'expected' to the statistical frequencies.

[b] See footnote c of Table I.

(c) Base composition and dinucleotides from intronic CRFs

As shown in Table I, GC levels of intronic CRFs range from 10% to 17%, with only one higher value, 25%, in the case of the CRF from intron rI. G is often in slight excess over C, A is usually more abundant than T. The GC level of intronic CRFs is two to three times higher than those of intergenic AT spacers (5%) and of the *var1* gene (8%).

Table II presents the analysis of AT dinucleotides for the CRFs of individual introns and for their totality (compiled end-to-end). In spite of their different GC levels, CRFs exhibit the same deviations (relative to frequencies expected for statistical sequences having the same size and the same GC level) previously detected in intergenic sequences (de Zamaroczy and Bernardi, 1987), except for the CRF of intron aI5β2, in which case deviations show opposite trends. With only two exceptions, AC and CA are present at lower levels than expected, as in the case of intergenic AT spacers.

The deviations just mentioned are (*i*) comparable with those seen in AT spacers, for CRFs from aI3, aI5β1, bI1, bI2 and bI3, which essentially are the introns of 'long' genomes; (*ii*) weaker, but still significant for CRFs from aI5α, bI4, bI5, aI6, which essentially are the introns present in supershort genomes; (*iii*) very weak (except for the excess of AT) for the two CRFs from rI. The compiled total CRF, which neglects CRFs of rI and aI5β2 (see above), exhibits the typical frequency deviations found in intergenic AT spacers of *S. cerevisiae*.

The higher levels of GG and CC are not comparable with those of intergenic sequences. Their lower excesses are

TABLE III

Copy number distribution of A/T tri- and tetranucleotides from the CRF sequences of mt introns of *Saccharomyces cerevisiae*[a]

3-mers	*A*	*B*	*A+B*
ATA	485	466	951
TAA	366	317	683
AAT	341	317	658
AAA	297	167	464
4-mers			
ATAT	243	236	479
ATTA	203	191	394
ATAA	206	180	386
AATA	191	188	379
TTAA	109	103	212
TAAA	127	77	204
AAAT	116	83	199
AAAA	128	68	196

[a] Column *A*: frequencies of A/T oligos. Column *B*: frequencies of complementary sequences. 'Expected' statistical frequencies are 631 ± 20 and 266 ± 20 for 3-mers and 4-mers, respectively.

largely due to the overall GC level of CRFs, which is due, in turn, to a much higher level of isolated G and C in CRFs, as well as to the contribution of the clusters of *n* group. The latter cannot be directly correlated with the typical intergenic GC clusters (de Zamaroczy and Bernardi, 1986a).

(d) Higher oligos from intronic CRFs

The results obtained with AT trinucleotides from the ensemble of CRFs indicate the same trends found in intergenic sequences. G/C trinucleotides are generally in excess, but less so than in intergenic sequences (not shown). Other trinucleotides are close to statistical expectations, except that trinucleotides with one C and with two C's are defective and in excess, respectively.

The analysis of AT tri-, tetra- and hexanucleotides in total CRFs shows frequency patterns comparable to those of AT spacers (Tables III and IV; Fig. 3). Indeed, (*i*) frequencies of complementary couples of isostichs decrease with decreasing *R* ratio (Fig. 3; *R* is the molar ratio AT+TA/AA+TT); (*ii*) isostich classes show an exponential distribution correlated with their *R* values, the slope in a semi-logarithmic plot being the same as in AT spacers; (*iii*) eight hexanucleotides (underlined isostichs in Fig. 3) exhibit the frequency of the immediately higher classes; seven of them are the same as those from the AT spacers; their rank in the decreasing order is almost identical to that from AT spacers (the only exception is AAAAAA); (*iv*) in some cases, complementary hexanucleotides do not exhibit the same frequencies, but their ratios are the same as in a statistical distribution; this is due to the fact that A and T are not equimolar in CRFs, whereas they are in AT spacers.

TABLE IV

Copy number of distribution of A/T hexanucleotides from the combined CRFs of mt introns of *Saccharomyces cerevisiae*[a]

6-mers	*A*	*B*	*A+B*
TATATA	83	82	165
TATTAT	72	66	138
ATTATT	63	57	120
TTATTA	57	55	112
TTATAT	52	42	94
ATTATA	47	45	92
ATATTA	49	41	90
ATATAA	45	43	88
ATTAAT	48	34	82
ATAAAT	44	28	72
AATTAT	36	32	68
TATTAA	38	26	64
TAAATA	36	25	61
AATAAA	40	21	61
AAATAA	36	19	55
TATAAA	34	20	54
AAATAT	28	26	54
AATATT	26	28	54
TTAATT	28	26	54
ATAAAA	34	16	50
AAAAAA	33	16	49
AAAATA	22	20	42
AAATTA	26	15	41
ATTAAA	21	18	39
TAAAAT	30	16	36
TAAAAA	24	11	35
TAAATT	18	13	31
TTTTTA	18	12	30
AATTTA	14	11	25
TTAAAA	14	10	24
AAAATT	15	9	24
AAATTT	7	6	13

[a] See footnote a in Table III. Hexanucleotides showing higher copy number than expected on the basis of their ratio are underlined (see section **d**).

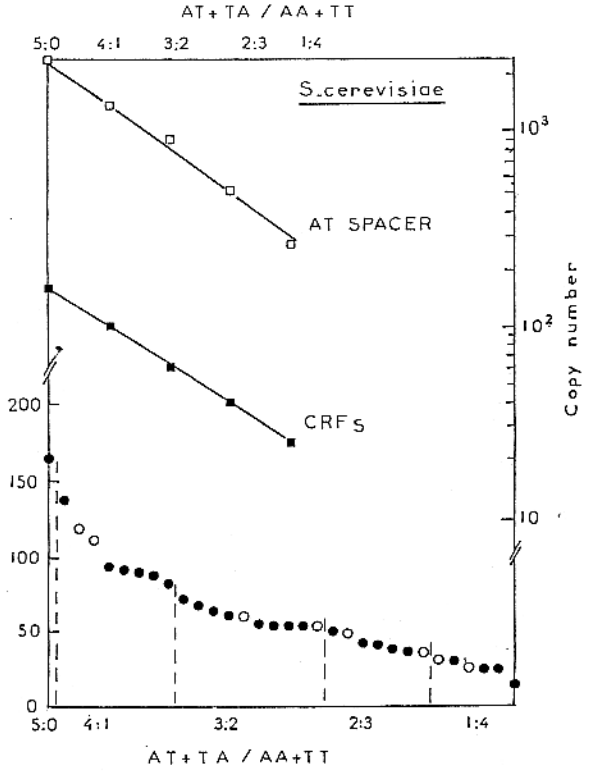

Fig. 3. Copy number distribution of AT hexanucleotides (see Table IV) from the total CRFs of mt introns of *S. cerevisiae*. Average copy numbers of AT hexanucleotides belonging to the same class are plotted against their *R* ratios (open and blackened squares refer to combined AT spacer and total CRF data, respectively). Blackened and open circles correspond to average values of pairs of complementary isostichs (6-mers) from the combined CRFs of *S. cerevisiae*. The corresponding *R* ratios are indicated on the abscissa. Open circles concern hexanucleotides which are more frequent than expected from their ratio. The data base used in the present work was GenBank Nucleotide Sequence Data Libraries under entry YSCMTCG, accession No. M62 622 (de Zamaroczy and Bernardi, 1986b). The computer programs used were either available at CITI2, Paris, or specially designed by C. Mugnier (random generation of sequences and isostich distribution).

TABLE V

(*A*) GC levels, base compositions, sizes (nt) and (*B*,*C*) frequency differences of AT di- and trinucleotides of four particular AT-rich segments from CRFs of mt introns of *Saccharomyces cerevisiae*

	Introns	G+C,%	A	T	G	C	Size (nt)
A	aI5β1	2.4	58.9	38.7	1.8	0.6	163
	bI3	2.2	51.7	46.1	0	2.2	89
	bI5/1	4.8	49.5	45.8	2.4	2.4	212
	bI5/2	4.4	46.5	49.1	2.6	1.8	114
	Total	3.6	51.9	44.5	1.9	1.7	578

		AA	TT	AT	TA
B	aI5β1	− 13.0	− 42.6	+ 24.6	+ 24.6
	bI3	− 27.8	− 41.2	+ 33.4	+ 33.4
	bI5/1	− 11.0	− 14.2	+ 15.1	+ 12.9
	bI5/2	− 18.1	− 8.3	+ 12.6	+ 12.6
	Total	− 15.0	− 22.7	+ 19.5	+ 19.5

C		AAA	TTT	ATA	TAT	ATT	TTA	AAT	TAA
	Total	− 14.3	− 35.2	+ 49.2	+ 62.1	− 19.4	− 10.7	− 15.0	− 15.0

(e) AT-rich sequences from intronic CRFs

Three CRFs of aI5β1, bI3 and bI5 belonging to the group-I introns (only present in 'long' genome, except for bI5), comprise four 120–220-nt-long AT-rich sequence elements that (*i*) are not involved in TSS; (*ii*) contain typical GC clusters; (*iii*) are present in loops other than the main ORF loop; and (*iv*) have a GC level very close to that of AT spacers. Table V shows that deviations of AT dinucleotides from these sequences are very similar to those exhibited by AT spacers. AT trinucleotides show a weaker under-representation and a stronger over-representation of non-alternating and alternating AT, respectively. Other AT trinucleotides are all under-represented compared to a statistical distribution.

(f) The sequences from intronic ORFs

Intronic ORFs (Table VI) generally show a number of quite different sequence features compared to intronic CRFs. GC levels are in the same range as those of CRFs in the case of ORFs from group I introns (11–20%), but they are higher in the case of ORFs from group II introns (25%). In all cases, AT dinucleotide frequencies are within 13% of the statistical expectation. Interestingly, however, AT dinucleotides show the same trends as those shown by intronic CRFs, although in a decreasingly pronounced way, in the case of ORFs aI5α, bI2, bI3 and aI3 (group I), but they do not in those of ORFs rI, aI5β, aI4, and bI4 (group I) and aI1, aI2 (group II). The most striking case is that of aI5α, in which base composition and AT dinucleotide patterns of both CRF and ORF are comparable to mt intergenic sequences. This is the more striking as aI5α encodes an endonuclease (Seraphin et al., 1992) and is mobile (like rI; Dujon, 1989).

(g) Sequence pattern comparison between intergenic and intronic ORFs

Three intergenic ORFs, ORF1, ORF2 and ORF4, also show some similarity with the first set of intronic ORFs (de Zamaróczy and Bernardi, 1987). More specifically, ORF4 is partially similar to ORF bI2, whereas ORF1 and ORF2 are similar to ORF aI3. In contrast, the other intergenic ORFs, ORF3 and ORF5, are definitely related to intergenic sequence (de Zamaroczy and Bernardi, 1987) and, as expected, do not show any close relatedness to intronic ORFs.

(h) Conclusions

The CRFs from mt introns of *S. cerevisiae* exhibit a very high similarity of isostich patterns with intergenic mt sequences. Indeed, (*i*) the deviations of the frequencies of dinucleotides from those expected for statistical sequences having the same base composition are very similar to those exhibited by AT spacers, although the GC level is considerably higher (by 8–12%; the maximum is 24% for aI3 and rI1) than that of intergenic sequences; (*ii*) copy numbers of AT tri-, tetra- and hexanucleotides have a ranking order very similar to that found in the same isostichs from the intergenic sequences; (*iii*) AT-rich segments from several intronic CRFs have base compositions and frequency deviations (from those statistically expected) of AT di- and trinucleotides, which are extremely close to those of intergenic sequences; (*iv*) GC clusters from intronic CRFs be-

TABLE VI

(*A*) GC levels, base compositions, sizes (nt) and (*B*) frequency differences of dinucleotides from ORFs of mt introns of *Saccharomyces cerevisiae*[a]

	Group I introns	G+C%	A	T	G	C	Size (nt)
A	aI5α	15.6	45.0	39.4	8.9	6.7	921
	bI2	17.0	44.1	38.9	9.7	7.3	846
	bI3	13.7	45.1	41.1	7.8	5.9	1048
	aI3	20.0	42.5	37.5	11.8	8.2	1005
	rI	20.0	42.2	37.7	11.4	8.6	708
	aI5β	14.7	43.7	41.6	8.3	6.4	954
	aI4	18.5	45.4	36.1	11.1	7.4	951
	bI4	19.5	42.4	38.1	11.4	8.1	1158
	Group II introns						
	aI1	24.7	42.8	32.4	14.6	10.1	2337
	aI2	26.4	38.8	34.8	15.8	10.6	2358

B	Introns	AA	TT	AT	TA	CC	GG	AC	CA
	aI5α	−7.4	−12.9	+11.9	+10.7	+140	+75	−30	−6.6
	bI2	−6.2	−6.0	+5.2	+12.2	+80	+133	0	−12.5
	bI3	−2.5	−7.1	+7.0	+8.1	+125	+83	−29.6	−25.9
	aI3	−1.7	−7.1	+3.8	+8.2	+14.3	+36	−14.3	−17.1
	rI	+7.3	+7.8	−5.0	−5.0	+129	+31	−22.2	−8.3
	aI5β	−1.1	+5.2	0	+0.6	+25	+86	−10.7	−7.1
	aI4	+2.9	−2.3	+6.7	−3.7	+33	+58	−41	+11.8
	bI4	+3.9	+1.4	0	+2.5	+57	+31	−11.8	−20.6
	aI1	−6.6	+2.9	+0.7	+7.9	+50	+33	+4.7	+9.3
	aI2	+2.0	+6.6	−0.7	+0.7	+91	+36	−12.2	−4.9
	Introns	AG	GA	TC	CT	TG	GT	CG	GC
	aI5α	+4.9	−7.6	+21.2	+2.3	−5.9	−0.2	—	−50
	bI2	+2.9	−15.8	−12.0	+16.2	−20.5	−12.5	−72	−1.1
	bI3	+2.3	−3.4	+7.2	+19.6	−6.4	−12.7	−57	+30.4
	aI3	+1.7	−8.3	+4.1	+20.3	−7.3	−2.8	−17	+13.7
	rI	+6.0	−2.3	−4.4	+4.9	−2.3	−6.9	−59	+12.2
	aI5β	+13.0	+10.0	−9.9	+16.4	−18.9	−39	−81	+69
	aI4	−6.7	−6.7	+16.0	+1.1	+12.3	−20.1	—	+70
	bI4	−8.9	−8.9	−19.0	+10.1	−1.0	−12.5	+8.3	+62
	aI1	+15.2	−2.4	+17.5	−2.2	−17.6	−4.9	−53	−11.9
	aI2	+4.4	−3.8	−5.1	−2.4	−10.9	−9.1	−34	−16.4

[a] See footnote a of Table II.

long in the same classes which are present in intergenic sequences.

The above similarities obviously point to an origin of intronic CRFs (with the exception of aI5β2 and of the CRFs from rI) from intergenic sequences. The differences indicate, however, a number of changes, which have led to an increase in the GC level of intronic CRFs relative to intergenic sequences. Apparently, a number of CRF stretches have still kept the original features of intergenic sequences. This may be conceivably due to the fact that their translocation was a recent one, and/or was compatible with the regional structures (see below).

As far as CRF formation and their association with intronic ORFs are concerned, the following remarks may be made. (*i*) If one takes into consideration the fact that CRFs and TSS cores of the original introns form inseparable ensembles (the deletion of CRFs leading to a defect in RNA-catalyzed splicing), one can suggest that insertions of short AT spacer sequences in the original intron did not disturb the TSS core; such sequences possibly even un-

derwent an expansion in situ. Alternatively, the presence of AT-spacer sequences (5% GC) in CRFs of aI5β1, bI3 and bI5 probably indicates more recent translocations of blocks of intergenic sequences, also hinted by the presence of typical GC clusters, which are all located outside sequences involved in TSS core. These sequences are not at all necessary for the functional structure of introns since they are not involved in TSS; (*ii*) most CRFs are downstream from the ORFs. This may be needed in order to avoid the interruption of the coding frame which exists between the upstream exon and the following intronic ORF. The purpose would be to protect the synthesis of chimeric (exon-intronic ORF) proteins that favor splicing (Burke, 1988).

In contrast with the case of most intronic CRFs, our analysis shows that the similarity of intergenic sequences to intronic ORFs is weak (except for aI5α) in the case of some intronic ORFs, or cannot be detected for the majority of them. Our analysis does not allow us to reach any decision concerning the origin of these ORFs. Yet it is certainly compatible with the possibility of horizontal interspecific transfer of introns (Lemieux and Lee, 1987; Dujon, 1989; Muscarella and Vogt, 1989; Bell-Pedersen et al., 1990; Séraphin et al., 1992). This hypothesis is so much more acceptable as the property of simultaneous handling of the splicing function and the capacity of self-transposition is known for certain introns (Wenzlau et al., 1989; Hickson, 1989; Perlman and Butow, 1989). Moreover, the existence of a significant similarity, both in primary structure and in the precise localization, has been shown for introns which belong to species that are very distant in evolution (Lang, 1984; Michel and Dujon, 1986; Zimmer et al., 1987).

Concerning the weak similarity noticed in the case of some intronic ORFs, we suggest that the initial ORFs incorporated CRF sequences by subsequent extension of an ORF (a 'sequence assimilation' similar to that described in the case of the *var1* gene; de Zamaroczy and Bernardi, 1987) and/or that the compositional constraints typical of the mt genome of yeast may have appeared in the case of ORFs which have been localized in the mt genome for a long time, another kind of 'sequence assimilation'.

Finally, in the case of intergenic ORFs 1, 2 and 4, which no longer have TSS, one can take into consideration a mechanism of translocation from the intronic ORFs (or from whole introns) towards intergenic regions, where TSS could disappear during evolution since the splicing function is not needed anymore .

To sum up, the present results demonstrate a mosaic organization in the mt introns of yeast! The evolutionary formation of CRFs appears to be the result of translocations of AT spacers (possibly followed by in situ sequence expansions) and of GC cluster insertions. In contrast, the formation of the ORFs is likely to be the result of events in which exogenous sequences capable of self-splicing have been transferred to mt genes. Indications of 'sequence assimilation' of ORFs and CRFs were found in some cases.

REFERENCES

Bell-Pedersen, D., Quick, S., Clyman, J. and Belfort, M.: Intron mobility in phage T4 is dependent upon a distinctive class of endonucleases and independent of DNA sequences encoding the intron core: mechanistic and evolutionary implications. Nucleic Acids Res. 18 (1990) 3763–3770.

Bernardi, G., Ehrlich, S.D. and Thiery, J.P.: Deoxyribonucleases: specificity and use in nucleotide sequence studies. Nature New Biol. 246 (1973a) 36–40.

Bernardi, G., Ehrlich, S.D. and Thiery, J.P.: A new approach to the study of nucleotide sequences in DNAs. In: Hamkalo, B.A. and Papaconstantinou, J. (Eds.), Molecular Cytogenetics. Plenum Press, New York, 1973b, pp. 95–99.

Bonitz, S.G., Coruzzi, G., Thalenfeld, B.E. and Tzagoloff, A.: Assembly of the mitochondrial membrane system. Structure and nucleotide sequence of the gene coding for subunit 1 of yeast cytochrome oxidase. J. Biol. Chem. 255 (1980) 11927–11941.

Burke, J.M.: Molecular genetics of group I introns: RNA structures and protein factors required for splicing — a review. Gene 73 (1988) 273–294.

Cech, T.R.: Conserved sequences and structures of group I introns: building an active site for RNA catalysis — a review. Gene 73 (1988) 259–271.

Cech, T.R.: Self-splicing of group I introns. Annu. Rev. Biochem. 59 (1990) 543–568.

Cech, T.R., Zaug, A.J. and Grabowski, P.J.: In vitro splicing of the ribosomal RNA precursor of *Tetrahymena:* involvement of a guanosine nucleotide in the excision of the intervening sequence. Cell 27 (1981) 487–496.

Cech, T.R., Tanner, N.K., Tinoco Jr., I., Weir, B.R., Zucker, M. and Perlman, P.S.: Secondary structure of the Tetrahymena ribosomal RNA intervening sequence: structural homology with fungal mitochondrial intervening sequences. Proc. Natl. Acad. Sci. USA 80 (1983) 3903–3907.

Davies, R.W., Waring, R.B., Ray, J.A., Brown, T.A. and Scazzochio, C.: Making ends meet: a model for RNA splicing in fungal mitochondria. Nature 300 (1982) 719–724.

Dujon, B.: Group I introns as mobile genetic elements: facts and mechanistic speculations — a review. Gene 82 (1989) 91–114.

Hickson, R.E.: Self-splicing introns as a source for transposable genetic elements. J. Theor. Biol. 141 (1989) 1–10.

Lang, B.F.: The mitochondrial genome of the fission yeast *Schizosaccharomyces pombe*: highly homologus introns are inserted at the same position of the otherwise less conserved *cox1* genes in *Schizosaccharomyces pombe* and *Aspergillus nidulans*. EMBO J. 3 (1984) 2129–2136.

Lemieux, C. and Lee, R.W.: Nonreciprocal recombination between alleles of the chloroplast 23S rRNA gene in interspecific chlamydomonas crosses. Proc. Natl. Acad. Sci. USA 84 (1987) 4166–4170.

Michel, F. and Dujon, B.: Conservation of RNA secondary structure in two introns families including mitochondrial-, chloroplast- and nuclear-encoded members. EMBO J. 2 (1983) 33–38.

Michel, F. and Dujon, B.: Genetic exchange between bacteriophage T4 and filamentous fungi? Cell 46 (1986) 323.

Michel, F. and Lang, B.F.: Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. Nature 316 (1985) 641–643.

Michel, F., Jacquier, A. and Dujon, B.: Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. Biochimie 64 (1982) 867–881.

Michel, F., Umesono, K. and Ozeki, H.: Comparative and functional anatomy of group II catalytic introns — a review. Gene 82 (1989) 5–30.

Muscarella, D.E. and Vogt, V.M.: A mobile group I intron in the nuclear rDNA of *Physarum polycephalum*. Cell 56 (1989) 443–454.

Nobrega, F.G. and Tzagoloff, A.: Assembly of the mitochondrial membrane system. DNA sequence and organization of the cytochrome *b* gene in *Saccharomyces cerevisiae* D273–10B. J. Biol. Chem. 255 (1980) 9828–9837.

Perlman, P.S. and Butow, R.A.: Mobile introns and intron-encoded proteins. Sciences 246 (1989) 1106–1109.

Pesole, G., Prunella, N., Liuni, S., Attimonelli, M. and Saccone, C.: Wordup: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucleic Acids Res. 20 (1992) 2871–2875.

Pietrokovski, S., Hirshon, J. and Trifonov, E.N.: Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. J. Biomol. Struct. Dyn. 7 (1990) 1251–1268.

Pizzi, E., Attimonelli, M., Liuni, S., Frontali, C. and Saccone, C.: A simple method for global sequence comparison. Nucleic Acids Res. 20 (1991) 131–136.

Séraphin, B., Faye, G., Hatat, D. and Jacq, C.: The yeast mitochondrial intron al5α: associated endonuclease activity and in vivo mobility. Gene 113 (1992) 1–8.

Van der Veen, R., Arnberg, A.C., Van der Horst, G., Bonen, L., Tabak, H.F. and Grivell, L.A.: Excised group II introns of yeast mitochondria are lariats and can be formed by self-splicing in vitro. Cell 44 (1986) 225–234.

Waring, R.B. and Davies, R.W.: Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing — a review. Gene 28 (1984) 277–291.

Wenzlau, B.M., Saldanha, R.J., Butow, R.A. and Perlman, P.S.: A latent intron-encoded maturase is also an endonuclease needed for intron mobility. Cell 56 (1989) 421–430.

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. and Bernardi, G.: The AT spacers and the *var1* genes from the mitochondrial genomes of *Saccharomyces cerevisiae* and *Torulopsis glabrata*: evolutionary origin and mechanism of formation. Gene 54 (1987) 1–22.

Zimmer, M., Wesler, F., Oraler, G. and Wolf, K.: Distribution of mitochondrial introns in the species *Schizosaccharomyces pombe* and the origin of the group II intron in the gene encoding apocytochrome *b*.Curr. Genet. 12 (1987) 329–336.