Gene, 41 (1986) 1–22
Elsevier

GENE 1494

The GC clusters of the mitochondrial genome of yeast and their evolutionary origin

(*Saccharomyces cerevisiae*; primary structure; noncoding sequences; *ori* sequences)

Miklos de Zamaroczy and Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jaques Monod, Tour 43 – 2, Place Jussieu, 75005 Paris (France) Tel. (1)4329-58-24; (1)4336-25-25, ext. 41.01

(Received August 5th, 1985)
(Revision received and accepted October 15th, 1985)

SUMMARY

We have studied the primary and secondary structures, the location and the orientation of the 196 GC clusters present in the 90% of the mitochondrial genome of *Saccharomyces cerevisiae* which have already been sequenced. The vast majority of GC clusters is located in intergenic sequences (including *ori* sequences, intergenic open reading frames and the gene *var1* which probably arose from an intergenic spacer) and in intronic closed reading frames (CRF's); most of them can be folded into stem-and-loop systems; both orientations are equally frequent.

The primary structures of GC clusters permit to group them into eight families, seven of which are related to the family formed by clusters A, B and C of the *ori* sequences. On the basis of the present work, we propose that the latter derive from a primitive *ori* sequence (probably made of only a monomeric cluster C and its flanking sequences *r** and *r*) through (i) a series of duplication inversions generating clusters A and B; and (ii) an expansion process producing the AT stretches of *ori* sequences. Most GC clusters apparently originated from primary clusters also derived from the primitive *ori* sequence in the course of its evolution towards the present *ori* sequences. Finally, we propose that the function of GC clusters is predominantly, or entirely, associated with the structure and organization of the mitochondrial genome of yeast and, indirectly, with the regulation of its expression.

INTRODUCTION

Early work from our laboratory showed that the mitochondrial genome of wild-type *S. cerevisiae* is characterized by a very low G + C content (18%), and by a striking compositional compartmentalization, about half of it being formed by long AT spacers made up of short (dAT:dAT) and (dA:dT) sequences, the other half covering a very broad range of G + C levels (Bernardi et al., 1968; 1970; 1972; Bernardi and Timasheff, 1970; Piperno et al., 1972; Ehrlich et al., 1972). Micrococcal nuclease degradation showed that AT spacers made up over 50% of the genome and had a G + C content lower than 5%, and that 10% of the genome had a G + C level as high as 65% (Prunell and Bernardi, 1974). Use of the restriction enzymes *Hae*III and *Hpa*II (splitting the sequences GGCC and CCGG, respectively) revealed (Prunell and Bernardi, 1977)

Abbreviations: bp, base pair(s); CRF, closed reading frame; nt, nucleotide(s); ORF, open reading frame; *ori*, origin of DNA replication.

TABLE I

List and location of GC clusters[a,b,c]

No.	Family	Location		Strain	No.	Family	Location		Strain	No.	Family	Location		Strain
1	*a3*	+ 4	*21S*	I	54*	*a2*	(322)	*9S*	M	107*	B			
2	*a3*	− 258	*thr2*	D1	55	*a1*	+ 74	*9S*	M	108*	A		*ori7*	B
3	*a2*	+ 367	*his*	D1	56	*n*	− 122	*pro*	M	109	*c*	+ 238	*ori7*	B
4*	*a2*	− 308	*leu*	D1	57	*n*	− 98	*pro*	M	110	*v*	(58)	ORF5	B
5*	*a2*	− 141	*leu*	D1	58/	*a4*	+ 115	*pro*	M	111	*a1*	+ 26	ORF5	B
6	*a2*	+ 110	*gln*	D1	59	*a4*	− 236	*ori1*	A	112	*a3*	+ 127	ORF5	B
7	*a4*	− 320	*lys*	D1	60	A				113	*a4*	+ 260	ORF5	B
8/	?	− 210	*lys*	D1	61	B		*ori1*	A;J	114	*c*	− 591	*ori2*	B
9*	*c*	− 170	*lys*	D1	62	C				115	*c*	− 552	*ori2*	B
10/*	*a1*?	− 26	*lys*	D1	63	*a1*	+ 225	*ori1*	A	116*	C			
11	*c*	+ 18	*arg1*	D1	(64)*	*a1*	− 909	*ori1**	B	117*	B		*ori2*	B
12/	*a1*?	+ 391	*gly*	D1	65/	*a1*?	−1600	*15S*	B	118*	A			
13/	*a2*	− 301	*asp*	D1	66/	*a1*?	− 900	*15S*	B	119/	*a1*?	+ 527	*ori2*	B
14	*a3*	+ 267	*asp*	A	67	*a1*	− 400	*15S*	B	120*	*a3*	+1000	*ori2*	B
15*	*a1*	− 109	*ser2*	D1	68	*a3*	− 355	*15S*	M	121	*a1*	+1100	*ori2*	B
16*/	*a1*?	− 355	*ala*	S	69	*a3*	− 296	*15S*	M	122	*n*	+1200	*ori2*	B
17	*n*	+ 129	*ile*	S	(70)*	*a1*	(603)	*15S*	M	123*	*a3*	+1500	*ori2*	B
18	*a3*	+ 222	*ile*	S	71*	*a1*	+ 138	*15S*	A	124	*a2*	+1600	*ori2*	B
19/	?	+ 401	*tyr*	S	72	*a2*	(125)	ORF3	A	125*	*a2*	+1700	*ori2*	B
20	*a1*	− 217	*asn*	D1	73	*a3*	+ 109	ORF3	A	126*	*a2*	+1900	*ori2*	B
21	*a4*	− 82	*asn*	D1	74	*a2*	+ 177	ORF3	A	127*	*a1*	+1950	*ori2*	B
22	*a4*	+ 48	*met*	D1	75	*n*	+ 271	ORF3	A	128	*c*	+2000	*ori2*	B
23	*a3*	+ 85	*met*	D1	76	*a2*	− 268	*trp*	A	129	*n*	+2100	*ori2*	B
24/	*c*	+ 411	*met*	D1	77	*a3*	− 40	*trp*	A	130*	*a4*	+2150	*ori2*	B
25*	*a3*	− 237	*oxi1*	D1	78	*a2*	+ 348	*trp*	A	131*	*a1*	+2600	*ori2*	B
26*	*c*	− 122	*oxi1*	D1	79*	*a1*	+ 609	*trp*	A	132*	*a3*	− 207	*glu*	A;D1!
27	*a1*	(864)	ORF1	D1	80	A				133/*	*a1*?	− 674	*cob*	D1;A
28/	*a2*	+ 344	ORF1	D1	81	B		*ori8*	B	134	*n*	− 368	*cob*	A;D1
29/	?	+ 230	*phe*	D1	82	C				(135)	*n*	(378)	bI1	F
30*	*a4*	+ 333	*val*	D2	83	*a4* (γ)				(136)	*n*	(717)	bI1	F
31*	*a1*	+ 189	*oxi2*	D2	84/	?	−1100	*oxi3*	K	(137)	*n*	(11)	bI2	F
32	*a4*	(197)	ORF2	D2	85/	?	− 300	*oxi3*	K	138	*a4*	(94)	bI5	D1
33	*a4*	(888)	ORF2	D2	86	*a4*	(26)	aI1	D2	139*	*a4*	(272)	bI5	D1
34	*a4*	+ 153	ORF2	D2	87	*a2*	(1)	aI2	D2	140*	*a1*	(573)	bI5	D1
35	*a4*	+ 323	ORF2	D2	88	*c*	(14)	aI3	D2	141*	*a1*	+ 372	*cob*	D1
36/	*a1*?	+ 447	ORF2	D2	89	*n*	(94)	aI3	D2	142	*c*	+ 502	*cob*	D1
37	*a2*	− 953	*ori5*	A	90	*n*	(140)	aI3	D2	143*	*a3*	+ 604	*cob*	D1
38	*c*	− 844	*ori5*	A	91	*v*	(174)	aI3	D2	144	*c*	− 573	*ori6*	D1
39	*a4*	− 644	*ori5*	A	92	*c*	(270)	aI3	D2	145	*a4*	− 457	*ori6*	D1
40	*a1*	− 574	*ori5*	A	(93)	*n*	(143)	aI5α	K	146*	*a4* (γ)			
41*	*a3*	− 386	*ori5*	A;K	(94)	*n*	(160)	aI5β	K	147*	C			
42*	C				(95)	*n*	(210)	aI5β	K	148*	*a2* (β)		*ori6*	B;D1
43*	B		*ori5*	A;K	(96)	*a1*	(127)	aI5β	K	149*	B			
44*	A				97	*a4*	− 274	*aapl*	J	150*	A			
45*	*a2*	+ 256	*ori5*	A	98*	*a2*	− 325	*oli2*	J;D2	151	*a1*	+ 19	*ori6*	D1
46*	*a1*	+ 437	*ori5*	A	99	*a3*	− 187	*oli2*	J;D2	152	*a4*	+ 249	*ori6*	D1
47/	*a2*	− 995	*fmet*	M	(100)*	*a1*	+ 393	*oli2*	D2	153	*a3*	+ 263	*ori6*	D1
48	*c*	− 957	*fmet*	M	(101)*	*a2*	+ 646	*oli2*	D2	154	*a4*	+ 565	*ori6*	D1
49	*a4*	− 878	*fmet*	M	(102)*	*a1*	−1127	*oli2*	D2	155/*	*a1*?	− 387	*oli1*	D1
50*	*a4*	− 728	*fmet*	M	103*	*a1*	− 371	*ori7*	B	156	*a1*	+ 600	*oli1*	D1
51*	*a4*	− 420	*fmet*	M	104*	*a2*	− 95	*ori7*	D2	157/	*v*	− 481	*ser1*	D1
52	*a3*	+ 19	*fmet*	M	105*	*a4* (γ)				158*	*a3*	− 193	*ser1*	D1
53	*v*	− 78	*9S*	M	106*	C		*ori7*	B	159*	*a1*	− 58	*ser1*	D1

No.	Family	Location		Strain	No.	Family	Location		Strain	No.	Family	Location		Strain
160*	*a2*	+ 227	*ser1*	D1	171/	?	− 250	*ori3*	A	182	A			
161	*v*	(187)	*var1*	D1;E	172	A				183	B			
162*	*a1*	+ 52	*var1*	D1;A	173	B		*ori3*	A;K	184	*a2* (*β*)		*ori4*	A
163	*v*	+ 420	*var1*	D1	174	C				185	C			
164	*c*	− 1800	*ori3*	A	175	*a4*	+ 137	*ori3*	A;K	186	*a4* (*γ*)			
165*	*a1*	− 1750	*ori3*	A	176	*n*	+ 432	*ori3*	A	187	*a1*	+ 200	*ori4*	A
166*	*a4*	− 1300	*ori3*	A	177*	*a2*	+ 637	*ori3*	A	188	*a3*	− 837	*21S*	I
167	*a4*	− 1050	*ori3*	A	178*	*a3*	+ 707	*ori3*	A	189	*n*	− 536	*21S*	I
168*	*a2*	− 900	*ori3*	A	179	*v*	− 429	*ori4*	A	190*	*a4*	− 414	*21S*	I
169*	*a3*	− 800	*ori3*	A	180	*a2*	− 315	*ori4*	A	191*	*a3*	− 136	*21S*	I
170*	*a1*	− 700	*ori3*	A	181	*a1*	− 159	*ori4*	A	(192)	*a4*	− 224	rI	I
										193	*n*	(6)	rI	I
										194	*n*	(205)	rI	I

[a] GC cluster numbering is clockwise on the genome map, with the *Sal*I site taken as the map origin (see Fig. 1). GC clusters, as presented in the figures, have a clockwise (no special indication) or anticlockwise (asterisks) orientation on the map; in other words, the former set of clusters are presented in the figures as read on the non-transcribed strand, the latter on the transcribed strand. Slashes after cluster numbers indicate incomplete primary structures or GC clusters only detected by diagnostic restriction sites. The nomenclature of GC-cluster families is given in RESULTS section; question marks indicate GC clusters which cannot, or can only tentatively, be classified. Cluster numbers in parentheses refer to clusters which are absent in some strains. Locations upstream (+) or downstream (−) of the nearest end of the genome landmarks (genes, *ori* sequences, intergenic ORFs and intronic CRFs; for this purpose, *ori* sequences are considered to extend from cluster A to cluster C only) are given in bp; distances in parentheses concern GC clusters located inside such landmarks; they separate the 5′ end of the cluster and the upstream end (in the clockwise sense) of the landmark; (in the case of introns the CRFs were taken as landmarks, except for clusters 96 and 137, for which the landmarks were the ORFs). Locations of landmarks on the genome map and all references concerning the primary structures are given by de Zamaroczy and Bernardi (1985). Some GC cluster locations are only approximate (distance ending with 00) owing to the presence of intra- or inter-regional gaps on the genome map (see Table II of de Zamaroczy and Bernardi, 1985). Yeast strains in which GC-cluster sequences were established are indicated with the single capital letter convention used in our laboratory (see de Zamaroczy and Bernardi, 1985): A, D243-2B-RI; B, C982-19d; D1, D273-10B/A21; D2, D273-10B/A48; E, A10; F, 777-3A; I, IL8-8C; J, J69-1B; K, KL14-4A; M, MH41-7B; S, SM202; it should be noted that strains A and J have common origin (Faugeron-Fonty et al., 1984) and that strains D1 and D2 are closely related. Exclamation marks indicate clusters largely different in primary structure in two different strains.

[b] Special cases. (i) GC-cluster pairs 22, 23; 68, 69; 152, 153 are contiguous and form G + C-rich stretches longer than 160 bp comprising three perfect direct repeats 104 bp long; (ii) GC clusters 71, 79 are identical over more than 110 bp; GC clusters 120 and 132 and their downstream flanking A + T-rich sequences are highly homologous over more than 100 bp in strains A and B, but different in strains A and D1; (iii) *ori*1 and its flanking sequences are different in strains A, J, and in strains B, C, K (de Zamaroczy et al., 1984; Faugeron-Fonty et al., 1984); in the case of strain B, sequence *ori*1* is characterized by a duplication of the whole cluster A–cluster B region and by an insertion beginning with a 96-bp GC cluster in sequence *1* and formed by partial clusters *β*, A, C, A and a whole cluster *β*; this insert which may be equated with two contiguous fused *a3* and *β* clusters is not listed in the table; (iv) *ori*8 is different in strains B, C, and in strains A, K (Faugeron-Fonty et al., 1984); *ori*4 is absent in strain B (Faugeron-Fonty et al., 1984); these findings imply interstrain differences in the corresponding GC clusters. (v) GC clusters 52–55 are inside the tRNA synthesis locus (Miller et al., 1983). (vi) GC clusters 100, 101, and 102 are inside ORF4 (Séraphin et al., 1985); these sequences fill in gap *11*.

[c] Clusters 137bis and 137ter are two contiguous clusters having an overall length greater than 160 bp and belonging to the *a3* family; they are located at positions 250 and 186, respectively, of CRF from b13, which was recently sequenced by R. Schweyen (personal commun.) in strain K; this sequence fills in gap *14*.

the existence of about 60 sequences in which *Hae*III and *Hpa*II sites were clustered together; these clusters were estimated to be 35 bp long on the average and to be 45–62% in GC. In addition, 40 isolated or clustered *Hpa*II sites were also found in the genome. Since the oligonucleotides released by *Hpa*II and *Hae*III only corresponded to 3% of mitochondrial DNA, other G + C-rich clusters, not containing *Hae*III or *Hpa*II sites, had to be present to account for the 10% of the genome having a 65% G + C level, and had to be larger in size and/or number than the *Hae*III-*Hpa*II clusters.

The proximity of restriction sites in at least a large fraction of GC clusters indicated that these were characterized by a certain degree of sequence homology and by potential palindromic secondary

structures. Such symmetry properties suggested by themselves a binding role for proteins having a dyad axis of symmetry (Bernardi, 1965; 1968) and a possible role as promoters, operators and initiation sites for DNA replication (Prunell and Bernardi, 1977). While the latter prediction and its corollary suggestion of multiple replication origins on the mitochondrial genome turned out to be correct (de Zamaroczy et al., 1981), the first was only correct in suggesting the presence of multiple promoters, since GC clusters do not seem to be involved in any direct way in the initiation of transcription. In fact, the idea that GC clusters were playing a role in the initiation of DNA transcription was largely based on the erroneous hypotheses (i) that the 40 isolated or clustered *Hpa*II sites were isolated sites located in tRNA genes, leaving a plausible number (60) of *Hpa*II-*Hae*III clusters as putative promoters of mitochondrial genes; (ii) that GC clusters containing and not containing *Hpa*II and *Hae*III restriction sites were clustered together; and (iii) that such GC clusters were located at AT spacer-gene borders. The latter point was disproven by sequencing work (Cosson and Tzagoloff, 1979; Gaillard and Bernardi, 1979), which showed that GC clusters were in fact embedded in AT spacers. Subsequent primary structure investigations (Tzagoloff et al., 1979; 1980; Berlani et al., 1980; Nobrega and Tzagoloff, 1980; Novitski et al., 1983) confirmed the symmetry and homology properties predicted for GC clusters (Prunell and Bernardi, 1977). Comparative studies of the primary and secondary structures of GC clusters were reported previously, but they were limited to small sets of GC clusters (Sor and Fukuhara, 1982; Goursot et al., 1982; Séraphin et al., 1985).

Here we have examined all the GC clusters which we have found in the available genome sequences; we have identified a relatively small number of families of GC clusters and we have reached a number of conclusions about the evolutionary origin and the functional role of GC clusters.

RESULTS

(a) The number and location of GC clusters

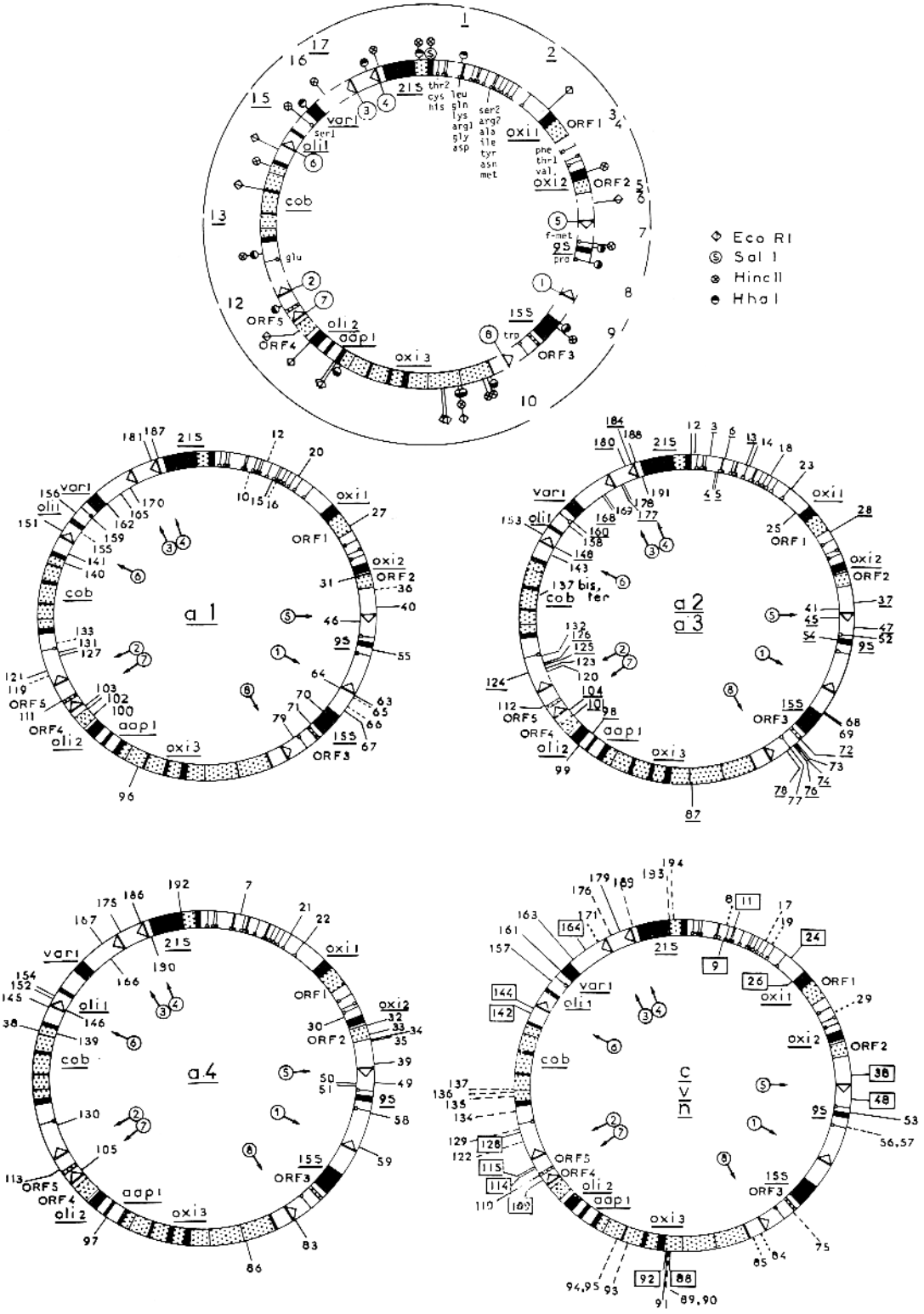
We have chosen to consider as GC clusters sequences (i) which comprise at least one G/C tetranucleotide, namely tetranucleotides only formed by G and/or C (e.g.: GCCG, CCGC etc.), and (ii) which are located in intergenic sequences (including *ori* sequences, intergenic ORFs and the gene *var1*), and in intronic CRFs. This operational definition will be justified in the DISCUSSION, section **d**, where other clusters, not fulfilling it, will also be discussed.

Table I presents a list of the 196 GC clusters as defined above, which we could find in all available sequences from the mitochondrial genome of yeast (see de Zamaroczy and Bernardi, 1985, for a review). GC clusters are numbered in a clockwise direction on the genome map, the *Sal*I site being taken as the map origin (Fig. 1). Their approximate locations are indicated in Fig. 1; precise locations are given in Table I relative to the nearest end of map landmarks (genes, *ori* sequences etc.). Orientations of GC clusters are given in both Table I and Fig. 1.

(b) The GC clusters A, B, C of *ori* sequences

24 GC clusters of the mitochondrial genome of wild-type yeast are found, as clusters A, B and C, in its eight *ori* sequences (Fig. 2a), the locations and orientations of which are shown in Fig. 1. These

Fig. 1. Physical and genetic maps of the mitochondrial genome of yeast showing the location of GC clusters. (Top circle) Physical map ▶ of a long mitochondrial genome unit of wild-type *S. cerevisiae* showing the remaining sequence gaps (de Zamaroczy and Bernardi, 1985); these are numbered 1 through 17 on the outside circle (gaps *11* and *14* have been just filled in; see footnotes to Table I); underlined numbers refer to intra-regional gaps, other numbers to inter-regional gaps (the former are gaps present within sequenced regions; the latter separate independently sequenced regions). In the sequenced regions, black areas correspond to mitochondrial genes or their exons; dotted areas to intervening sequences and intergenic ORFs (ORF1–ORF5; Colin et al., 1985); radial lines indicate tRNA genes. Among mitochondrial genes, *oxi*1, 2 and 3 encode subunits II, III and I, respectively, of cytochrome *c* oxidase; *cob*, cytochrome *b*; *aap1*, *oli2* and *oli1*, subunits 8, 6 and 9 of ATPase; *var1*, a protein associated with the small sub-unit of mitochondrial ribosomes; *9S* corresponds to the tRNA synthesis locus; *15S* and *21S* are the genes for the small and large ribosomal RNAs, respectively. Circled numbers indicate the location of cluster C of *ori* sequences; arrowheads point in the direction cluster C to cluster A. Restriction sites are those present in the mitochondrial genome units of strain A (de Zamaroczy et al., 1984). (Other circles) Approximate locations of GC-cluster families



on the genome map. Precise locations are given in Table I. GC-cluster numbering is that of Table I. When the 5′–3′ orientation of the GC clusters, as given in the figures, is clockwise, numbers are outside the circle; when anticlockwise, they are inside. (*a1* family) Broken lines refer to GC clusters only putatively assigned to the family. (*a2* and *a3* families) Underlined numbers refer to the *a2* family. (*c, v, n* families) Boxed numbers refer to the *c* family; other numbers to the *v* family (solid lines), to the *n* family (broken lines) and to GC clusters unknown in their primary structure (dotted lines).

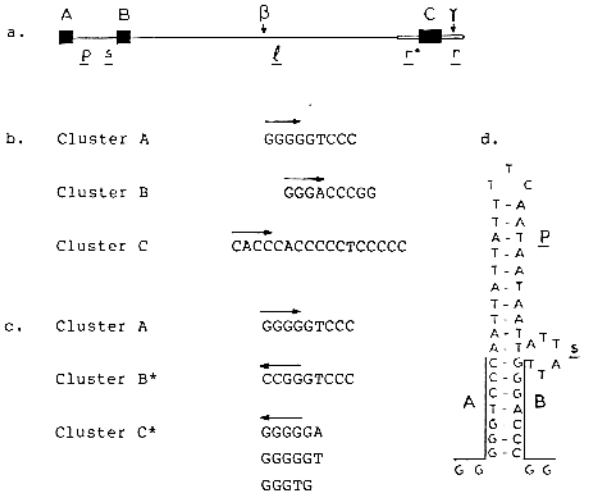


Fig. 2. The *ori* sequences of the mitochondrial genome. (a) A schematic representation of one of the *ori* sequences. Clusters A and B are separated by a palindromic AT sequence, *p*, and by a short AT sequence, *s*. Cluster C is flanked by two short AT sequences *r** and *r*, where RNA-primed bidirectional DNA replication starts. Clusters B and C are separated by the long AT sequence *l*. The position of clusters *β* and *γ*, which are only present in some *ori* sequences, are indicated. (b) The GC clusters A, B and C from the *ori* sequences of the mitochondrial genome of yeast. The primary structure is read on the strand carrying the oligo(C) stretches of cluster C. The (GGT) sequence precedes cluster A in three inactive *ori* sequences. (c) The complementary inverted sequences of the B and C clusters are compared with the A cluster sequence. (d) Potential secondary structure of clusters A and B and of the intervening palindromic AT sequence *p* and short AT side-loop *s*.

clusters are closely related in primary structure (Fig. 2b). Cluster C is made up by three penta-C sequences (one of which contains an A) separated by an A and a T, respectively (Fig. 2b). The penta-C motif preceded by an A is the inverted complement of six of the nine bp of cluster A (Fig. 2c). Cluster B (Fig. 2b) is largely (seven out of nine bp) the inverted complement of cluster A (Fig. 2c), with which it can form (Fig. 2d) a palindromic structure comprising the intervening AT sequences, *p* and *s* (Gaillard and Bernardi, 1979; de Zamaroczy et al., 1981).

(c) The other GC clusters

Characteristic primary structure features allowed us to group these GC clusters into seven families which are all related to the GC clusters of *ori* sequences. Interestingly, the identification of these families is not easily done by computer-assisted comparisons, because of base changes, insertions and deletions, and also because of sequence mistakes.

The properties of each family are described below. The first GC cluster family, *a1* (previously referred to as *ori*[s]; Goursot et al., 1982), shows the highest degree of homology among its members. The subsequent three families, *a2–a4*, show both a decreasing homology among their members and also a decreasing homology with *a1*. As far as the other three families are concerned, *v* and *c* show a certain degree of homology with the *a2* and the C clusters,

TABLE II

Location, number and relative amounts of GC clusters in the mitochondrial genome of yeast[a]

Family	Intergenic sequences			Intronic		Genes				Number	Amount %
	ori	ORFs	Others	CRFs	ORFs	*15S*	*21S*	*var1*	ORF5		
a1		1 + (2)	34	1	(1)	(1)				40	24
a2	2	1 + (1)	22	1						27	15
a3			24	(2)						26	26
a4	4	2	22	3			(1)			32	17
v			4	1				1	1	7	4
c			13	2						15	5
n			15	3 + (6)	(1)					25	5
A,B,C	24									24	4
Total	30	7	134	19	2	1	1	1	1	196	100

[a] Parentheses indicate facultative GC clusters, or GC clusters present in facultative introns; facultative clusters or introns are absent in certain strains.

respectively, whereas the last one, *n*, groups short GC clusters which do not fall into any of the other families.

Table II summarizes the location, number, and relative amount of the clusters belonging to the different families.

(d) The *a1* family

As just indicated, this family of GC clusters is characterized by the highest degree of homology among its members (Fig. 3). Indeed, the homology of the 31 *a1* clusters with their consensus (majority) sequence is comprised between 90 and 100%. The length of the *a1* clusters is 44 bp, its GC content 70%. Deletions are rare among family members and concern 1–3 bp, except for three cases (clusters 63, 140 and 156), where a longer deletion seems to be present in the central part of the sequence. In all likelihood, this is due to the loss of short restriction fragments comprised between the two *Hpa*II or *Hae*III sites which were used for terminal labelling. As a consequence of the rarity of deletions and additions, and of their occasional compensation, the length of *a1* clusters is very constant, within 1–2 bp in most cases. Out of the 44 bp, 34 are essentially constant (13 positions are changed only once), and stretches 9–18 and 34–42 are the best conserved. Four of the ten variable positions (positions 7, 23, 43 and 44) are the most susceptible to deletions or substitutions.

Nine putative *a1* clusters are not completely known (clusters 10, 12, 16, 36, 119, 133, 155; see Fig. 3) or have only been mapped by restriction enzymes (clusters 65, 66; see Table I).

The *a1* clusters are characterized by the following sequence features: (i) an initial octanucleotide, *i*, CTCCTTTC; (ii) an octanucleotide, *a*, GGGGTTCC, which is identical with cluster A of *ori* sequences except for the lack of the first G and for a C → T change at the underlined position; (iii) a 13-bp stretch, *m*, GGCTCCCGTGGCC; (iv) a non-anucleotide, *b*, GGGCCCCGG, which is identical with cluster B except for the replacement of an *A* with a C, at the underlined position; and (v) a terminal hexanucleotide, *t*, AACTAT. Some degree of homology is also found in the flanking AT sequences; the majority sequences are TATTT on the *i* side and AAA on the *t* side.

GC clusters of the *a1* family can be folded into hairpin structures (Goursot et al., 1982), as shown in Fig. 4. Most changes occurring in the family affect nucleotides not involved in the base pairing of the stem structure.

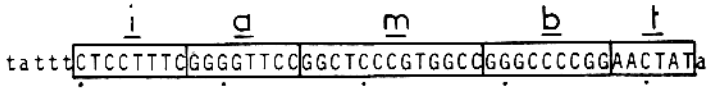
(e) The *a2* family

This family (Fig. 5) comprises 27 members having a size of 43 ± 10 bp and a GC content of 72 ± 7%, except for four shorter (28, 47, 124, 125) and two longer clusters (160, 168). They are characterized by the following features: (i) an initial sequence, *i*, CTCCTT–C, which differs from that of the *a1* family only in lacking the sub-terminal T; this sequence is regularly preceded by a TA doublet; (ii) sequence *a*, GGGGT–CC, which differs from cluster A of *ori* sequences in lacking a G in its first position and a C in the indicated position; (iii) a sequence, *m*, GCCCCGCGGGGGC, similar to sequence *m* of the *a1* family; (iv) a sequence, *b*, GGG––CCGG, differing from sequence B of *ori* sequences in lacking the central AC doublet; (v) a sequence, *t*, –ACTAT, identical to that of family *a1*, except for the lack of the first A.

The consensus sequence of the GC clusters of the *a2* family, as established on the basis of the sequence elements discussed above, can be folded into a hairpin structure (Fig. 4). The two identical *β* clusters (148, 184), inserted in *ori* 4 and 6, can also be folded into stem-and-loop structures (de Zamaroczy et al., 1984); these clusters are characterized by the partial fusion of *a* and *m* sequences.

(f) The *a3* family

This family (Fig. 6) is formed by 26 GC clusters ranging in size from 42 to 126 bp and having a GC content of 45–67%. Some of these clusters are arranged in tandem with clusters of the *a2* family. The primary structure of the members of this family is characterized: (i) by a sequence, *i*, identical in most cases with that of the *a2* family; this is preceded by an A and sometimes followed by a TTTAAAA bridge; (ii) by a diagnostic sequence *a–c* which is present in the majority of cases, and which exceptionally (cluster 132) may even precede sequence *i*; the *a–c* sequence is formed by (1) an *a* sequence, similar to that found in the *a1* family except for the lack of




```
 15*  ttttttatattattt-TCCTTTCGGGGT-CCGGCTCC-GTGGCCGGGCCCCGGAACTATtaataagtaataaat
 20            °°°t-TCCTT-CGGGGTTCCGGCTCCCGTGGCCGGGC T CCGGAACTATtaataaaattaataa
 27   ctaattataatatttCTCCTTTCGGGGTTCCGG-TCCCGTGGCCGGGCCCCGGAACTA-aaaatattattgatg
 31*  tctatatgtatatttCTCCTT A CGGGGTTCCGGCTCCCGTGGCCGGGCCCC A GAACTA T aaaaataaaattatt
 40   tattataatttccttCTCCTTTCGGGGTTCCGGCTCCCGT A GCCGGGCCCCGGAACT T Taaaagaaaaaaggga
 46*           °°°CTCCTT-CGGGGTTCCGG A CCCGTGGCCGGGCCCCGGAACTATtaattaatataggtg
 55   attaaaatatttattCTCCTTT A GGGGTTCCGGCTCCCGTGGCCGG AA CCCGGAACTATtaattaataataaat
 63   ataaaaaatatatttCTCCTT A CGGGGTTCCGGCTCC-GT A GC-GGGCC°°° A GAACTA A ataaaatatattatt
 64*  tatatatataaataaCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTA T aaatattatatatta
 67   aatggtttaataataCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTAT t aataataaataata
 70*  aaatatttaatatttCTCCTTTCGGG-TTC-GGCTCC-GTGGCCGGGCCCCGGAACTA-aataaaataaaatat
 71*  tttatatttatatttCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACT--CCTCCCTTGCCTTGC
 79*  attattgtataatttCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACT--CCTCCCTTGCCTTGC
 96   attttcaattaatatCTCCTTT T GGGGTTCCGG-TCCC-TGG C CGG C CCCCG-A AACTA-aagatattaagaatt
100*  ataataatttataatCTCCTT G CGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTATaatatttttgaatt
102*  aatttatttttaatatCTCCTT GT GGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTATaatatttttaatagg
103*  aatatgttataatttCTCCTT C CGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTATatgatataataaata
111   aataatttattaattCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGG AA C CCGGAACTATaaaaataatttttaat
121   taatataaacttctcCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTATtaattatataatata
127*  attatattattatttCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTATtataattatatatat
131*  atatttatcttattcCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTATtaatatataaattga
140*  atataatatatattaCTCCTTTCGGGGTTCCGGCTCCCG°G°°°°°°°°CCGG T ACTATataaaaaaaaattaa
141*  taataattaataaatCTCCTT G CGGGGTTCCGGCT T CCGTG A CCGGGCCCCGGAACTATaatatataaaataat
151   caaatatatatataaCTCCTT-CGG G GTTCCGG T TCCC G TGGCCGGG G CCCGG A ACTA T aatatattatatata
156   atattggtgaaacatCTCCTTTCGGGGTTCCGGCTCCCG°G°°°°°°°°CCGGAACTATtaatatttaataaaa
159*  taaatattaaataaaCTCCTTTCGGGGTTCCGGCTCCCGTGGCCG AA CCCCGGAACTATaaatatattatatat
162*  atatttatttttatttCTCCTTTCGGGGTTCCGG-TCC-GTGGCCGGGCCCCGGA-CTATtaatatatatattaa
165*  ttattaaatatccttCTCCTTTCGGGGTTCCGGCTCC-GT A GCCGGGCCCCGGAACT T Taaaataaaggaataa
170*  aatatttatatacttCTCCTTTCGGGGTTCCGGCTCCCGTGGC-GGGCCCCGGAAC-ATtaagattataacaag
181   tattaatttttatttCTCCTT C CGGGGTTCCGGCTCCCGTGGCCGGGCCC-GGAACTATtaaatataaatatat
187   aaataattatactttCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACT T Tataatattattatta

 10*  acataatatatattaCTCCTTT°GGGG°°CCGG°°°°°°°°°°°°°°°°CCGG°ACTATtattaatatatatat
 12   ataatttataataatCTCCTT G T°GGGG°°°°°°°°°°°°°°GGCCGG°°°CCGG°°CTATaatatatttttaata
 16*  atattttatattattCTCCTTTCGGGGTTCCGG°°°°°°°GGCC°°°°°CCGG°°CTATtaaatattttattta
 36   attattaatattattCTCCTTTCGGGG°°°°°°°°°°°°°GGCCGG°°°°°°°°°°°°°
119   aaataaataataata-TCCTTT T GGGGTTCCGG°°°°°°°°°°°°°°°°°°°°°°°°°°
133*  ttattaatttatattCTCCTTTCGGGGTTCCGG°°°°°°°°°°°°°°°°CCGGAACTATaaaaaaaaaaataaaa
155*  tatatataaataaatCTCC°°°°°°°°°°°°°°°°°°°°°°°°GGCCGG°°°°°GGAACTATaattttatttattac


 /71*  GGGATTGGTTCACACCTTTATAATTAAATAAAGGTGTTCACTATAAATATT
 /79*  GGGATTGGTTCACACCTTTA------AATAAAGGTGTTCACTAT T AATATT
```

Fig. 3. Primary structure of the GC clusters from the *a1* family. The high-homology sequences are written in capital letters, the low- or no-homology flanking sequences in lower-case letters. The top line represents the consensus sequence (boxes corresponding to the sequence elements are discussed in RESULTS, section **d**) and to the low-homology flanking sequences. Dots indicate positions 1, 11, ▶

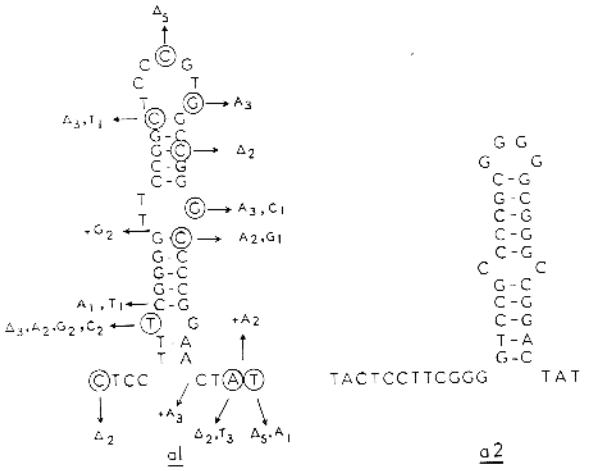


Fig. 4. Putative secondary structure of the GC clusters from the *a1* and *a2* families. Circles indicate variable positions. Base changes, deletions and additions occurring more than once in the family *a1* are indicated by the replacing base, by the symbol *Δ*, or by the inserted base preceded by a + sign; subscripts indicating the number of changes. In the case of sequences 71 and 79, two additional stem-and-loop systems can be formed downstream of the *a1* loop, the 6-bp deletion of sequence 79 corresponding to a loop (not shown). A more complex secondary structure can be formed for the region comprising cluster 71 and the sequence immediately following the *15S* gene (R. Martin, personal commun.).

the last C; (2) a sequence *c* identical to 9 bp of the C cluster of *ori* sequences; and (3) an intermediary GGT sequence (which forms a partial *a* sequence together with the first three nucleotides of the *c* sequence); sequence *a–c* is sometimes followed by GGGA, or AC and/or TA(A)GTATA; (this latter motif, *q*, may be repeated, up to five times in cluster 112); (iii) by an *a* sequence, identical to cluster A, but lacking in most cases the first G; sequence *a* is followed, (1a) in about half of the cases, by another *i* sequence, CTCCTTC, preceded by TCA; (1b) in other cases, by a GC cluster reminiscent of the *a2* family, (clusters 23, 69, 137ter, 153); (2) by a partial *a2* cluster (cluster 14) and an atypical G + C-rich sequence (clusters 18, 112); (3a) by an atypical G + C-rich sequence (cluster 1) which may comprise a modified *i* sequence at the junction (clusters 25, 191); or (3b) by a second cluster reminiscent of the *a3* family following a modified *i* sequence (cluster 77).

A putative secondary structure can be formed by base-pairing sequence *a* with a part of sequence *a–c*, sequence *q*, forming a loop (not shown). The putative secondary structure of clusters 68–69 (Fig. 7) is characterized by a series of five hairpins; the same structure can be drawn for clusters 22–23 and 152–153, except for the first stem-and-loop structure, which is variable, and for some changes in the last one.

(g) The *a4* family

The 32 GC clusters of this family are related to those of the *a1–a3* families in that both, or (more rarely) one of their end elements are similar to sequences *i* and *t*, respectively; these sequences may also be located inside the clusters; sequence *t* is preceded by a GG dinucleotide corresponding to the end of sequence *b* from families *a1* and *a2* (Fig. 8). About half of the *a4* clusters contain a 10–22 bp sequence corresponding to the beginning or the central part of the *a1* consensus sequence, or possess an *a* sequence (of the *a1* type) at different locations.

The clusters of the *a4* family basically differ from those of the other families described so far in that (i) their central part is variable in sequence and length; clusters 7 and 21 only consist of sequences *i* and *t*; clusters 32, 33 and 34 are *a1* clusters with a central deletion of about 15 bp. (ii) they generally have a GC content of 50 to 70%; and (iii) they cannot be folded into hairpin structures, except for the four *γ* clusters (de Zamaroczy et al., 1984) of the *ori* sequences (83, 105, 146, 186) and for the mini-insert (192) of the *21S* RNA gene (Dujon, 1980).

(h) The *v* family

The seven GC clusters of this family (Fig. 9) are related to a cluster present in the *var1* gene

21, 31, and 41 in the *a1* sequence (at the top of the figure). Insertions are indicated by subscripts, substitutions by superscripts, small deletions by dashes, sequence gaps (possibly corresponding to lost restriction fragments in the case of sequences 63, 140, 156) by small circles. Incompletely known *a1* clusters are listed separately (cluster 10, etc). The homologous right flanking sequences of clusters 71 and 79 are shown on the two bottom lines. Asterisks indicate sequences having an anti-clockwise orientation on the map (see footnotes to Table I). In two cases (clusters 71 and 79), *a1* sequences are followed by a homology region extending over 66 bp, which is rich in G + C in its first half and very rich in A + T in its second half. As a consequence, the regions comprising these clusters form two direct repeats 112 bp long, which only differ by 1-bp substitution and a 6-bp deletion (Bordonné, 1982; Martin et al., 1983).

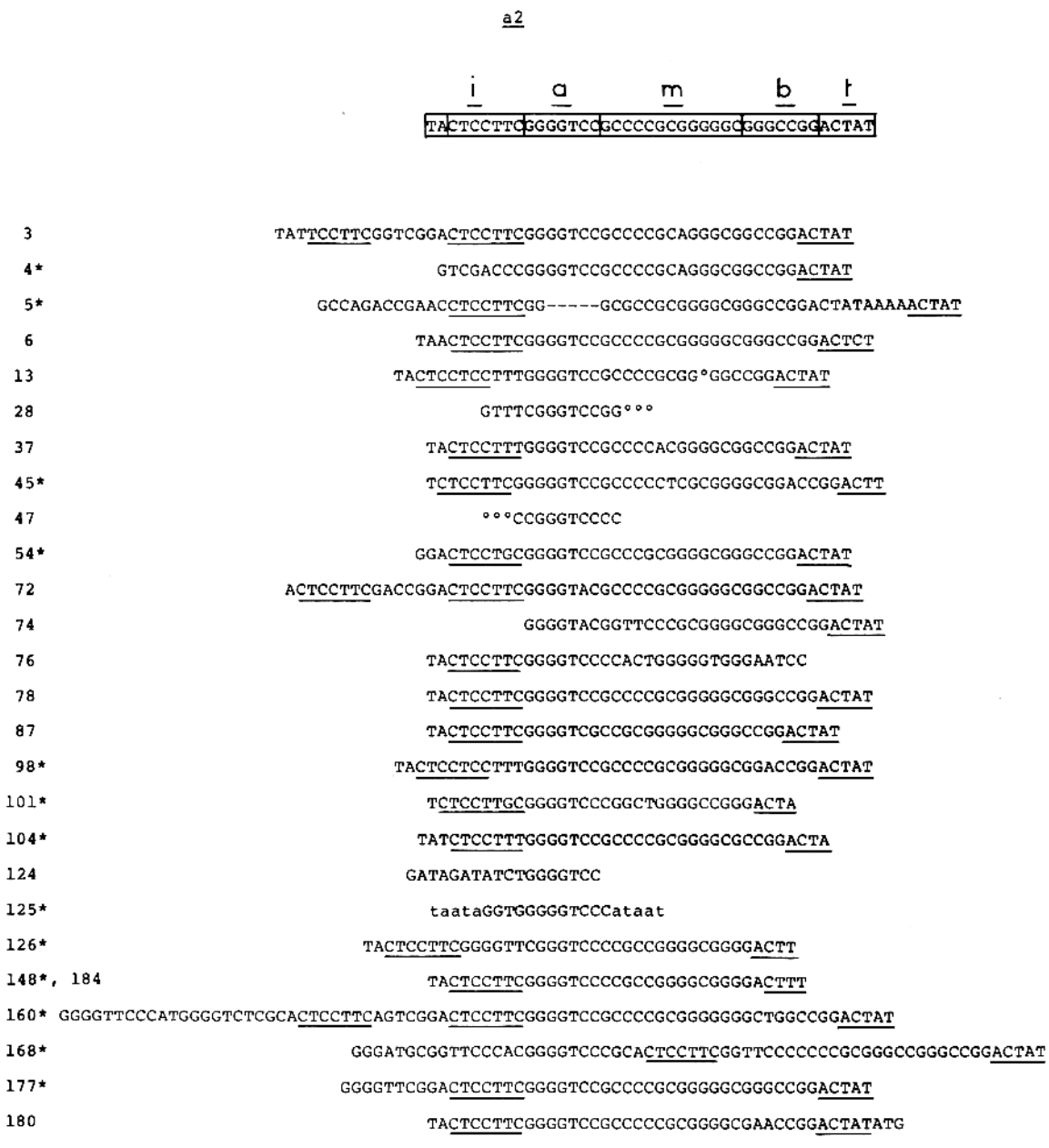


Fig. 5. Primary structure of the GC clusters from the *a2* family. The top line presents the consensus sequence, boxes correspond to the sequence elements discussed in RESULTS, section e; *i* and *t* sequences are underlined in individual sequences. Small circles indicate sequence gaps. In two cases (clusters 5 and 177) members of the *a2* family are associated with atypical G + C-rich sequences located upstream of sequence *i*, or (cluster 76) replacing the stretch following sequence *a*. In some cases, sequence *i* is preceded by an atypical G + C-rich sequence which comprises a duplicated *i* sequence (clusters 3, 72, 160), or is replaced by some elements from sequence *a*–*c* of the *a3* family (clusters 4, 168). The four shorter *a2* sequences only show sequence *a*; in one case (cluster 125), this sequence is a perfect cluster A.

(Hudspeth et al., 1982). They are very rich in G + C, 68–89% (with one lower value, 48% for cluster 110), have a size of 34–48 bp (with one 18-bp long exception), lack (like the *c* and *n* families) sequences *i* and *t*, are often flanked by 2–14 bp inverted or complementary inverted repeats only formed by A and T, share the same orientation and can be folded into hairpin structures with an identical terminal loop in at least five cases (Fig. 10). The *v* clusters are identical (in two cases), or very similar in primary structure. Two exceptions are cluster 110, located in ORF5, which only shares the primary structure of

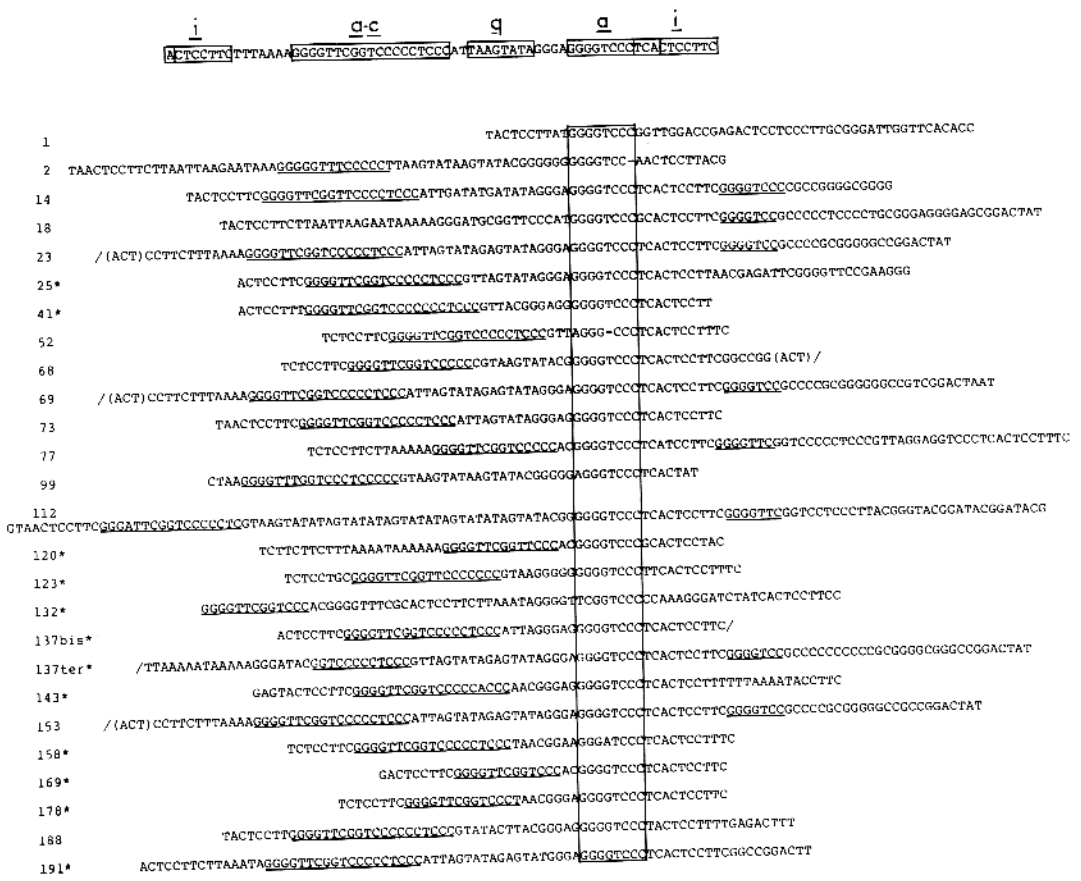


Fig. 6. Primary structure of the GC clusters from the *a3* family. The top line presents the consensus sequence, boxes correspond to the sequence elements discussed in RESULTS, section f; sequences *a–c* and *a'* are underlined. Slashes indicate that the cluster is preceded or followed by another cluster; in these cases the initial or terminal ACT, in parentheses, is shown in both clusters. The modified *i* sequence of cluster 68 is followed by partial *b* and *t* sequences GGCCGGACT (as described for the *a2* family); this terminal ACT sequence is also part of the *i* sequence of cluster 69, which is formed by an *a3* cluster followed by an *a2* cluster lacking sequence *i*. Clusters 68 and 69 form together a GC cluster of 164 bp, the longest found so far in the genome. Likewise, clusters 23 and 153 immediately follow clusters 22 and 152 (two clusters of the *a4* family), respectively, and comprise an identical 104-bp motif also present in clusters 68–69. This homology (which can reach 117 bp, if 2 or 3 point mutations are neglected) defines the three longest common motifs present in the mitochondrial genome (except for the *ori* sequences and for clusters 71 and 79 of the *a1* family). Clusters 137bis and 137ter are contiguous.

the loop and the secondary structure with the other members of the family, and cluster 163, which corresponds to a partial *v* cluster. The clusters of this family are different from those of the other families but exhibit a GGTCCCGCGGGG sequence corresponding to a fusion of *a* and *m* sequences from the *a2* family. A partial *m* sequence also precedes this motif. Three clusters (157, 161, 163) of this family are present inside, before and after the *var1* gene, respectively.

(i) The *c* family

This family is formed by 15 clusters which contain a 7–13-bp sequence identical to a part of cluster C of *ori* sequences (Fig. 9). They are short, 8–27 bp, are different from each other and from other GC clusters, cannot be folded into hairpin structures and predominantly show the orientation of cluster C of *ori*1 (Fig. 1). Cluster 114 differs from cluster C only by two changes and the lack of the two first bp. Cluster 92 is much longer than the others (58 bp) and contains a rare nona-G sequence which can base-pair with the cluster C-like sequence; (another nona-G is present in clusters 2 and 123, and a nona-C is present in cluster 4 and 137ter). At variance with the other families of GC clusters, the *c* family shows a strong asymmetry of C and G; the ratio C/G, which is about one for the other families (except for *a4* and

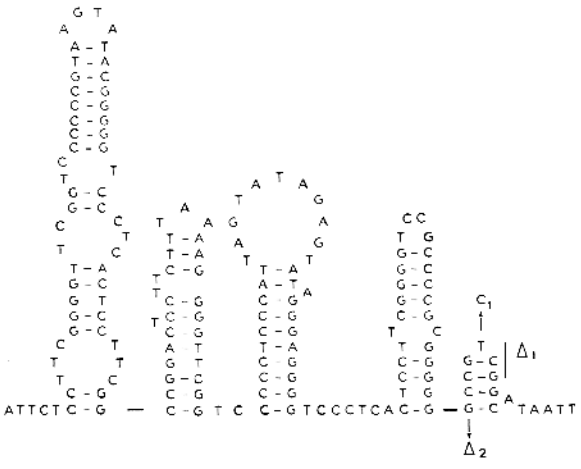


Fig. 7. Putative secondary structure of clusters 68–69 of the *a3* family. The same structure can be built for clusters 22–23 and 152–153, except for differences in the first (not shown) and last hairpin structures. Base changes and deletions are indicated as in Fig. 4.

n), is comprised between two and six; some clusters only contain C on one strand. The clusters of the *c* family are often present in pairs.

(j) The *n* group

This group comprises 19 mostly short and very short clusters (Fig. 9) which cannot be directly correlated with any of the families described so far and do not comprise in most cases G and/or C sequences longer than tetranucleotides; some of them contain short direct or complementary inverted repeats. Six additional GC clusters (8, 19, 29, 84, 85, 171; see Table I) cannot be classified at present because they are only known by restriction mapping.

(k) Restriction sites in GC clusters

Only a few restriction sites are found in GC clusters, the most frequent ones being *Hae*III and those cutting clusters A and B of *ori* sequences, namely *Hpa*II, *Ava*II (or *Sau*96I).

a4

```
  7    TACTCCTTTGGGACTTAT
 21    TCTCCTTCCGGAACTAAT
 22    TACTCCTCCTAGCAGGATTCACATCTCCTTCGGCCGG(ACT)/
 30*   TATCTCTTTCGCCGGATTAT
 32    ACTCCTTACGGGGGTTCCCGCGAAGCGGGAACTGGAT
 33    TCTACTCCTTACGGGGGTTCCCGCGAAGCGGGAACTA
 34    TATTCCTTCAAAACTCCTAACGGGGTTCCCGCGAAGCGGGAACTAAT
 35    TATTCCTTCGAGGTCACCGCCTCACCTCCAGCGGGACTTT
 39    GCGGACACCGTTACGCGAGTGGGGACTAT
 49    TAACCCCTTCC
 50*   TCCTTTGGGGAC
 51*   TATCCTTTGGGGTTCCCCCATGGG
 58    GAACCCTTCGGGGTTCGGT°°°
 59    TCTCCTTTCGGGGTTCCGGCTCCGGCCGGAAAATCCCTCCCAGTCCTGACGGGGCTTGCTCACATCCTT
 86    GGAAAGCCGTATGATGGGAAACTATCACGTACGGTTTGGGAAAGGCTC
 97    TATCTCCGCAAAGCCGGATTAATG
113    GAGTCCCTCACTCCTTATCACTNCGCTGAAGGTGG
130*   TCTCCTTTTGGGATTCCGGTCTTTGGCCTGGTATCC
138    TCTCCTTTCGGGGTT
139*   ACTCCTTATGGGAGTTCCCACAAAGCGGAAACTTAAG
145    GGGGGATCCCTCTCTCATCCGGCGTCCTACTCACCC
152    TCTCCTTCGACCGG(ACT)/
154    GGGATTCGGTTTCCCTCATCCTCATGGGTATCCCTCACTCCTTCTG
166*   GGGGAGACTCTTTCGGTGGTCCTGCCGCGGGCGGGCCGGACTAT
167    CCCCGATGGGGACTTAT
175    TACTCCATTAGGGGTTTTGGTCCATATCGGGAACCGAACTAT
190*   CTCATCTCCTTCGGTCGGACTAT
192    TCTCCTCTCTCGGTGGGGGTTCACACCTATTTTTAATAGGTGTGAACCCCTCTTCGGGGTTCCGGAACTT

83, 105*, 146*, 186   CGTCTCCGAGGTCCCGGTTTCGTAAGAAACCGGGACTAT
```

Fig. 8. Primary structure of the GC clusters from the *a*4 family. Sequences *i* and *t* (preceded by GG) discussed in RESULTS, section g, are underlined, the *a* sequence is boxed. For the explanation of slashes see Fig. 6.

v

```
 53  taataaaCCCCCCGCGGGCGCCAACCCCGTTGCTCACCGGGTTGGTCCCACGGGGttaaataat
 91  CCCCACGAGGGCCACACATGTGTGGCCTCGCGGGGTTTGG
110  aataCGGGTAAACATTACCCGTTGTTCACGGGTAATGTTTACCCtatt
157  CCCCGCGGGCGCCAAT°°°°°°°°°°°CCGGATTGGTCCCGCGGGG
161  aataataatataAACCCCGCGGGCGCCAATCCGGTTGTTCACCGGATTGGTCCCGCGGGGaatattaataataa
163  aaCCGCCCCCGCGGGCGTAGtt
179  attGGGCGCCAAGCCGGTTGTTCACCGACTTGGTCCCaat
```

c

```
 9*  CCACCCTCACTATCC
 11  CCCCCCCC
 24  CCCTCCCCCTTCCAAAATTCCGG°°°
 26* CCAAACCCCACC
 38  GGGGTTCGGTTCCCCCCTCACTCATTC
 48  GGCACCCACACACACAAAC
 88  CTTTCTTCCCCTCCGAATCCG
 92  CCCCACCCCCATGCGAAGCATGGGGGGGGGGTATAAGTATGGAC
109  CTGGGGACCACCCCCCCACC
114  CTCCTCCCCTCCCCCACAC
115  GGGCGCCAACCCCGCAGTG
128  GTCCGTGGGATCGAACCCCCT
142  GTCCCCGCCCCC
144  CTGGGGTTTCCCCCACTTAC
164  GTCTGCCCCCTCTTC
```

n

```
 17  GTTCCCGCGAAGCGGGAACCCCATAAGGAG
 56  CCCC
 57  GGGGACT
 75  GGGGAC
 89  GATGCCG
 90  GCGGGAAAACCCTAAAGTC
 93  GCGTAGAGATTAGACGCCTCTGGTTATCTAAG
 94  GCTAACGGGGTTTCTC
 95  CCCGTG
122  GATACGGATCAAATATTACCCGTTGTTCACTGGCAATG
129  GGGG
134  CCACGGGACCAATGACCAACCCAGTAGTTGACCGGATTGGCGCCCGCGAGG
135  CTTGGTAAGCCC
136  CGTACAGTTCTGAGTGGGGG
137  CGCCTC
176  CGAGGTATTAGTATCAGTATCAGTATCAGTATCGTAAAAAACGGGTGAC
189  GTTCACATTGGAGGCGAGTAAAAAGGAG
193  CCCCCTTGTCCC
194  CATCCACGAGCGCCCAATGTCG
```

Fig. 9. Primary structure of the GC clusters from the *v*, *c* and *n* families. Some sequence elements discussed in the RESULTS, sections **h** and **i**, are underlined. The flanking inverted or complementary inverted repeats of some *v* clusters (see Butow, 1985) are written in lower-case letters.

About 130 clusters comprise at least one *Sau*96I site, GGNCC, predominantly represented by GG$^{A}_{T}$CC, which corresponds to an *Ava*II site. About 100 clusters comprise at least one *Hpa*II site; 61 clusters comprise at least one *Hae*III site; except for three of them, they are all clustered with at least one *Hpa*II site; these are the *Hae*III-*Hpa*II site clusters originally recognized by Prunell and Bernardi (1977). Only seven *Hae*III sites and four *Hpa*I sites are not localized in the GC clusters of

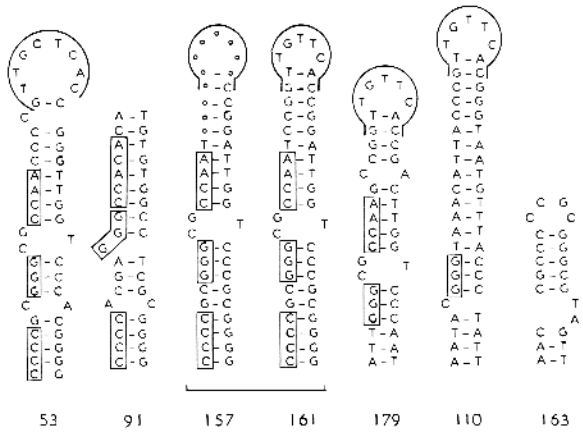


Fig. 10. Putative secondary structure of GC clusters from the *v* family. Homologous sequences are indicated by boxes or by circles for the loops (in the case of cluster 157 homology is based on indirect evidence). Numbers refer to Table I.

Table I; three *Hae*III and two *Hpa*II sites are located in the intronic ORFs of *oxi3*, two *Hae*III and two *Hpa*II sites in the intronic ORFs of *cob* and two *Hae*III sites in the *21S* gene. Finally, 30 clusters comprise one *Tha*I (CGCG) site and nine comprise one *Hha*I (GCGC) site.

Some restriction site associations characterize specific families of GC clusters: (i) A *Sau*96I site (not of the *Ava*II type) comprised between two *Hpa*II-*Hae*III sites characterizes family *a1*. (ii) One *Ava*II site and one *Hpa*II site exist in sequences *a* and *b*, respectively, of the *a2* family; in about half of the cases there are, in addition, a *Hae*III and/or a *Tha*I site. (iii) Two *Ava*II sites, localized in the *a* sequence and in the diagnostic sequence *a–c*, respectively, characterize family *a3*, which is alsmost completely deprived of *Hpa*II and *Hae*III sites. (iv) One *Hpa*II site is present just before sequence *t* of about half the GC clusters from the *a4* family. (v) The association *Hpa*II-*Ava*II is present in most *v* clusters, and is accompanied by a *Tha*I site in five cases and/or by a *Hha*I site in four cases. (vi) Practically no sites are found in families *c* and *n*, and in half of the *a4* clusters.

About 50% of GC clusters have no *Hpa*II or *Hae*III sites, and 30% also do not have *Sau*96I or *Hha*I or *Tha*I sites. Most of the latter clusters are shorter than 35 bp and comprise all of the shortest clusters.

DISCUSSION

(a) The evolutionary origin of *ori* sequences

It is reasonable to assume that the most ancient sequence elements in the *ori* sequences are cluster C (or rather its basic monomeric penta-C unit) and its contiguous sequences, *r** and *r*, since these are by far the most important elements in the *ori* function. Indeed, bidirectional DNA replication at *ori* sequences is initiated by RNA primers starting at sequences *r* and *r**, respectively, and continuing into DNA chains at cluster C (Baldacci et al., 1984). Moreover, cluster C is an evolutionarily highly con-

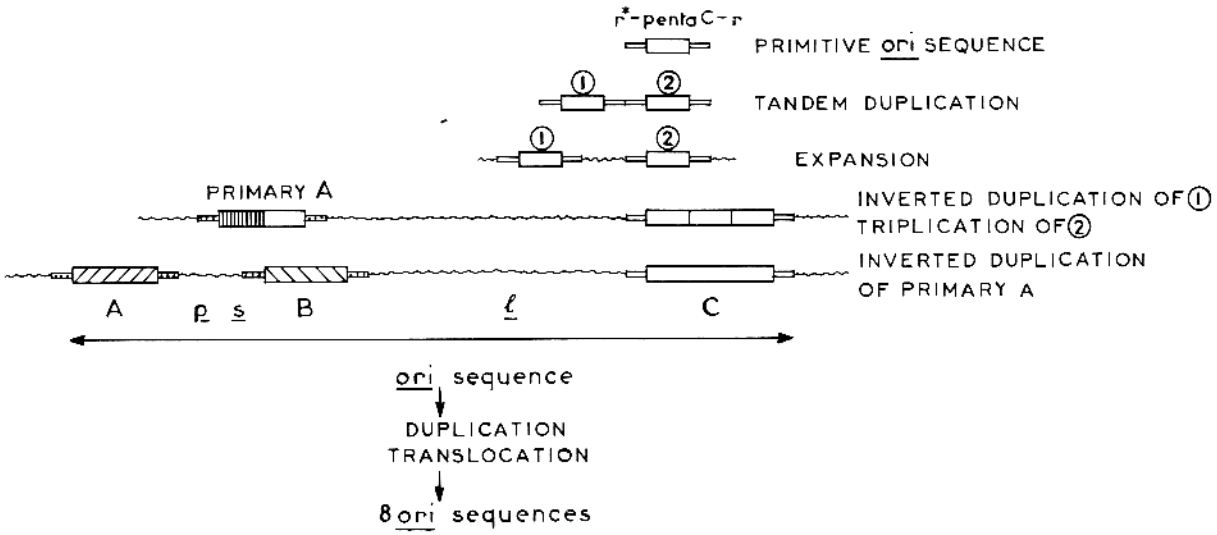


Fig. 11. Hypothetical scheme for the evolutionary construction of *ori* sequences. See DISCUSSION, section a, for details.

served sequence, which has been identified (de Zamaroczy et al., 1981; 1984) with sequence CSB2 (Wong et al., 1983) of vertebrate mitochondrial genomes.

The formation of an *ori* sequence might then have taken place through the steps presented in Fig. 11, namely (i) a duplication of the *r**-penta-C-*r* system (namely the postulated primitive *ori* sequence), followed by the functional inactivation of the duplicated sequence; (ii) the expansion, through a slippage mechanism accompanied and/or followed by internal duplication and/or intermolecular unequal crossing-overs, of the sequence separating the two penta-C systems; (iii) the triplication of the original penta-C sequence; (iv) the formation of cluster A by an inverted duplication of the duplicated penta-C; (v) the formation of the A-B region through the duplication-inversion of cluster A and its flanking AT region.

Needless to say, the only purpose of the scheme of Fig. 11 is to show (i) that simple duplication-inversion mechanisms (such as those demonstrated to be operational in the mitochondrial genome of yeast; de Zamaroczy et al., 1983; Faugeron-Fonty et al., 1983; Mangin et al., 1983), and expansion by slippage, duplication and recombination (as postulated for the formation of intergenic sequences and CRFs of intervening sequences; Bernardi, 1982; 1983) can account for the formation of the present *ori* sequences and (ii) that the formation of the first *ori* sequence, as described above, was followed by duplication and translocation events leading to the present eight *ori* sequences.

(b) The other GC clusters

All other GC clusters show more or less extensive homologies with each other and with the A, B, C clusters of *ori* sequences. These homologies are stressed by Fig. 12a, which points out the sequence elements shared by the clusters of the *ori* sequences and of the *a* families; these clusters represent together 86% of all clusters (see Table II).

Another point made by Fig. 12a concerns the sequences flanking the GC clusters of the *a1–a4* families; on the *i* side there is a small sequence which varies in size from one to five nt according to the family and shows a certain degree of conservation; on the *t* side, there is some conservation over three nucleotides only in the case of family *a1*. These findings raise two related questions (i) that of the definition of the borders of GC clusters of the *a1–a4* families; and (ii) that of the existence of "acceptor" sequences for cluster insertions. An attempt to answer the first question can be made by looking at the ends of GC clusters inserted in highly conserved regions of the mitochondrial genome, namely into the *ori* sequences and the *15S* and *21S* genes. The results (Table III) concern the seven clusters which can be investigated from this viewpoint. They clearly indicate (i) that sequence *i* is part of the inserted clusters; (ii) that the *i* end of the inserted sequence may vary by one nucleotide; (iii) that the few nucleotides preceding sequence *i* roughly match those found in the other sequences of the *a* families (Fig. 12a), in which a precise study of the ends could not be made.

The situation is less clear for the *t* end of the

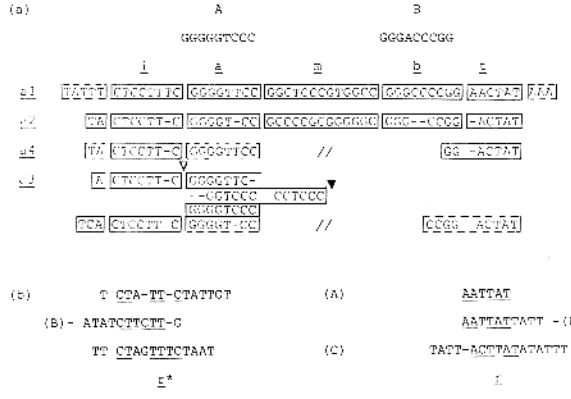


Fig. 12. Comparison of *ori* sequences and GC clusters. (a) A comparison of the primary structures of GC clusters from the *a1–a4* families. Common sequence elements are boxed (with dashed lines when sequences show more variability) and compared with GC clusters A and B from *ori* sequences. Double slashes stand for variable sequences; open and closed triangles indicate two inserts in clusters of the *a3* family: TTTAAAA and AT TAAGTATA GGGA/AC, respectively. (b) Sequences flanking clusters A, B and C of *ori* sequences are compared with sequences *i* and *t*. As far as the *i* homologous sequences are concerned (i) the sequence flanking cluster A is that of *ori*1, a lower homology is found for *ori*2 and no homology for the other *ori* sequences; (ii) the flanking sequence of cluster B follows instead of preceding the cluster; (iii) the flanking sequence (*r**) of cluster C is that of *ori*2, the other *ori* sequences showing lower similar homologies. The *t* homologous sequences are identical in all *ori* sequences, except for the insertion of clusters γ in *r* sequence of *ori*4, 6, 7, and 8, and for the fact that the sequence precedes instead of following cluster B. The correlation between the sequences flanking clusters B and the scheme of Fig. 11 is not obvious.

TABLE III

Insertion of GC clusters[a]

Location	Inserted cluster	Family
21S	TAACTTTCTCCTCTCTC.....CCGGAACTTA	*a4* (192)
15S	AATATTTCTCCTTTC.....CCGGAACTAAAT	*a1* (70)
*ori*4,6	GTTAATACTCCTTC.....GGGGA–CTT	*a2* (β) (184,148)
*ori*4,6,7	ATTACGTCTCC.....CCGGA–CTAT	*a4* (γ) (186,146,105)

[a] References are: Dujon (1980) (*21S*); Sor and Fukuhara (1982) (*15S*); de Zamaroczy et al. (1984) (*ori*). Inserted sequences of clusters (underlined) are defined by sequence comparison, except for the case of *15S* gene where S1 mapping was used; in this case the assessment of the cluster ends is only good within ± 2 nt.

clusters. Here, in the case of the only *a1* cluster of Table III, an almost complete *t* sequence appears to belong to the cluster and is followed by two to three A's as in the general case of Fig. 12a; in the case of the clusters from the other families, only one to three nucleotides from the *t* sequence are found in the inserted cluster and the flanking sequence may be either variable or, in the case of the γ clusters, correspond to the last three nucleotides of sequence *t*. It is obvious from the above that no clearcut conclusion can be drawn yet and that more precisely analyzed cases are needed. It appears, however, that on both sides for the clusters of the *a1* family and on the *i* side only for the others, the nucleotides flanking the inserted clusters are not random indicating some sequence specificity for insertion. It is of interest that the family, *a1*, which shows a higher internal homology, possibly reflecting the most recent series of insertion events, is also the one which shows more evident signs of non-randomness in the flanking sequences; this may indicate that the original "acceptor sequences" were blurred over evolutionary time by point mutations.

As far as the other families of GC clusters are concerned, they are all characterized by the lack of sequences *i* and *t*, a feature apparently related to their different evolutionary origin (see section **c**, below). It should be noted (Fig. 9) that inverted or complementary inverted repeats only formed by A- and T-flank clusters of the *v* family.

(c) The evolutionary origin of the other GC clusters

Several pathways seem to have been followed to generate the GC clusters other than A, B and C during evolution. In the case of the GC clusters of the *a1–a4* families, the pathway might be related to the evolutionary construction of *ori* sequences, as outlined in section *a* above: GC clusters from the primitive *ori* sequence evolving towards the present *ori* sequences might be duplicated and translocated with their flanking sequences, undergo duplication, inversion and/or recombination events similar to those described in Fig. 11. In connection with this suggestion, it should be mentioned that a certain degree of homology seems to be recognizable between sequences *r** and *r* on one hand, and sequences *i* and *t* on the other; the same seems to apply, to a lesser degree, to the sequences flanking clusters A and B (Fig. 12b); obviously, the last point is of interest in connection with the model proposed for the formation of *ori* sequences.

The primary clusters so originated were then amplified by duplications and translocations (and also by regional duplications; see section **e**, below). The amplification phenomenon is strongly indicated by (1) the homology of GC clusters belonging to the same families, the outstanding example being that of the *a1* family; such high homology is suggestive of bursts of amplification-insertion events which have been recent enough not to have been followed yet by a large divergence; alternatively, such GC clusters may have been immediately used for some function and protected against divergence (see, however, section **g**, below); (2) the identification of GC clusters, which are obviously inserted (see preceding section).

Other pathways were apparently followed to generate the GC clusters of the *v* family, of the *c* family, and of the *n* group. The clusters of

the *v* family are very similar to each other and their prototype might have arisen from unequal recombinations between two *a2* clusters leading to a motif related to the *a* and *m* sequences, GGTCCCGCGGGG (the underlined and overlined parts corresponding to parts of the *a* and *m* sequences, respectively) which can generate a *v* cluster by an inverted duplication of the *m* part. The *c* family is clearly related to cluster C from which its members might have derived. Finally, the clusters of group *n* might be the result, at least in part, of deletions and additions resulting from recombination events, operating on the other clusters described so far. In rare cases they might also be the result of statistical distributions of G's and C's (see section **d**, below).

(d) The definition of GC clusters

The operational definition of GC clusters used in the present work (see RESULTS, section **a**) is justified by the following considerations: (i) intergenic sequences (including *ori* sequences, intergenic ORFs, and the gene *var1*) have an overall size of 53 000 bp and a G + C level of about 15% (de Zamaroczy and Bernardi, 1985); intronic CRFs have an overall size of 5000 bp and a G + C level of 20%; (ii) statistical sequences having the size and G + C level of intergenic sequences only contain 7–10% of the G/C-tetranucleotides, which are present in the mitochondrial sequences just mentioned; in other words, the probability of these tetranucleotides arising just as a result of a statistical distribution of G's and C's is very weak; such a chance does, in fact, decrease even further if additional compositional constraints found in GC clusters (such as additional G's and C's contiguous or proximal to the tetranucleotides) are taken into account; for example, G/C-pentanucleotides, are found in the statistical control sequences at 0.5% the level at which they exist in the mitochondrial sequences mentioned above; incidentally such pentanucleotides are present in 83% of all GC clusters.

Some GC clusters were taken into consideration, in spite of the fact that they did not fit the operational definition because of their localization or of their primary structure, for the following reasons: (i) the two facultative clusters present in the *21S* and *15S* genes (Table II) have an obvious insertional origin (Table III); (ii) the two clusters present in the ORFs of the facultative introns from "long genomes" aI5β and bI2 (Table II) are located in sequences having a G + C level (17%) close to those of the contiguous CRFs (12–14%) and of intergenic sequences; comparable G + C levels also exist in intronic ORFs aI5α (16%) and bI3 (14%) which do not comprise any GC cluster; (iii) some clusters from the *a4* (7, 113, 154, 190) and *c* (9, 48) families are clearly related to other members of the same families.

It should be noted that no other clusters fitting the operational definition exist anywhere else in the genome, except for clusters of the *n* group located in genes and intronic ORFs; these were not taken into consideration, however, because of the high G + C levels of the harboring sequences. We also did not take into consideration about 60 G/C-trinucleotides located in intergenic sequences; in our opinion most of them should be considered as related to GC clusters, because, if we eliminate the latter (as defined above) from the intergenic sequences, the G + C level of such sequences drops to 5%, and statistical sequences having their overall size and the same G + C level only contain 10% of the trinucleotides found; moreover the trinucleotides actually found are only made up by G or C in 50% of the case, and if most trinucleotides are isolated, several are in the vicinity of dinucleotides only formed by G and/or C.

(e) The amount, number, distribution, location and orientation of GC clusters

The GC clusters as defined above represent 9% of the mitochondrial genome of yeast and about 13% of the intergenic sequences and of intronic CRFs. The total number of GC clusters found by us in 77 000 bp of long mitochondrial genomes is 196, including facultative clusters, only present in some strains; since the sequenced regions represent about 90% of the genome, one can assume that the total number of GC clusters is 210–220 per genome unit. This value accounts for the previous findings of 60 *Hae*III-*Hpa*II clusters, of 40 isolated or clustered *Hpa*II sites (which we now know to be almost always located outside genes; see RESULTS, section **k**), and of a larger amount of GC clusters not containing *Hpa*II or *Hae*III sites (Prunell and Bernardi, 1977).

The distribution of some GC clusters witnesses some past events in the evolutionary history of the mitochondrial genome of yeast. For example, (i) the

two groups of clusters between ORF3 and *ori8* and between *ori2* and *glu*, respectively, suggest that these regions arose by a duplication-inversion phenomenon; (ii) the identical sequence (over 100 bp) and orientation of clusters 71 and 79 (including about 30 bp of contiguous AT spacer) suggests a duplication and translocation of this region; likewise, the identical tandem sequences (over 100 bp long) carried by three pairs of clusters 22–23, 68–69, and 152–153, also speak for duplications and translocation. Similar conclusions were drawn previously for the duplication-translocation of *ori* sequences (de Zamaroczy et al., 1984).

As far as the distribution of GC clusters in different yeast strains is concerned, we have already stressed elsewhere the lines of evidence in favour of interstrain homology among intergenic sequences and we have pointed out that clusters located in genome regions which were sequenced in different strains are generally identical in primary structure (de Zamaroczy and Bernardi, 1985). This accounts for the compilation of GC clusters from different strains in a single table, Table I. The rare case in which large changes in sequence were found are in all likelihood due to site-specific unequal crossing-overs (see section **g**, below).

GC clusters located outside *ori* sequences are overwhelmingly present in intergenic sequences (Table II). The exceptions to the rule are seven clusters in intergenic ORFs 1–4, two clusters in two intronic ORFs, four clusters in mitochondrial genes (*15S, 21S, var1* and ORF5), and 19 clusters in intronic CRF's; interestingly, nine of them (which belong to the *n* group or can be assimilated to it, viz. clusters 86, 88, 90, 93, 94, 95, 136, 193 and 194), are comprised in the sequences involved in class 1 and class 2 structures (Michel and Dujon, 1983; Waring and Davies, 1984), whereas the other ten (eight of which do not belong to the *n* group) are not.

Concerning the latter clusters, it should be pointed out that while the GC clusters of the *15S* and *21S* genes are facultative inserts which are found in some strains only, those of *var1* and ORF5 reflect a different situation. Indeed, when the 1-kb sequence of *var1* was compared with 1 kb around *ori*1, it was found that these stretches have comparable high numbers of direct and inverted repeats and many common sequences (Bernardi and Bernardi, 1980). This finding is in favour of the idea that *var1* is a gene which has recently arisen from an intergenic sequence (Bernardi, 1983), a view supported by the fact that it is not even conserved in another ascomycete, *Neurospora crassa*. The evolutionary story of ORF5 is likely to be similar to that of *var1*.

As to the location of the other clusters, it should be noted (i) that the evidence for ORF1–4 being actual coding sequences is very weak and that, for the time being, these sequences are best considered as other intergenic sequences (Colin et al., 1985); see, however, Séraphin et al., 1985; (ii) that the finding of GC clusters in introns will be commented upon elsewhere.

Both orientations of GC clusters are almost equally represented on the mitochondrial genome except for the clusters of the *v* and *c* families (see Table I and Fig. 1). This finding is not surprising in view of what has been said on the evolutionary origin of clusters and of the rearrangements undergone by the evolutionary genome of yeast.

(f) GC clusters in the mitochondrial genomes of other unicellular eukaryotes

An important question is whether the GC clusters studied here are a peculiarity of the mitochondrial genome of *S. cerevisiae* or are also found in other genetic systems. A number of results are in favour of the second alternative. A first set of data concern three mitochondrial genomes, from *Euglena gracilis*, a green alga (Stutz and Bernardi, 1972; Fonty et al., 1975), *Ustilago cynodontis*, a basidiomycete, and *Acanthamoeba castellani*, a protozoan (Mery-Drugeon et al., 1981), and two chloroplast genomes, from *E. gracilis* and *Spinacia oleracea*, a plant (Schmitt et al., 1981). All these genomes are characterized by low G + C contents (ranging from 25% to 33%, with a higher value, 36.5% for *S. oleracea*) by the presence of A + T-rich stretches (25–30% of these genomes is made up of stretches as low as 12–15% G + C, with a higher value 22% for *S. oleracea*) and of G + C-rich clusters (at least 10% of the DNAs is higher than 60% in GC). A second case is that of the GC-repetitive sequences containing *Sac*II sites from *Kluyveromyces lactis* (Sor and Fukuhara, 1982).

More detailed information is available for the mitochondrial genome of *N. crassa*. In this case, ten clusters were identified in intergenic and intronic

sequences (Yin et al., 1981) and shown to be characterized by an 18-bp palindrome (the "core sequence") comprising two *Pst*I sites and forming the central part of a 50–100 bp cluster averaging 70% in GC. It should be noted that *Pst*I palindromes of *N. crassa* differ from GC clusters of *S. cerevisiae* in size and primary structure in showing one orientation only and a distribution clustered around rRNA and tRNA genes. The total number of these GC clusters containing *Pst*I sites is presumed to be 50 to 100 (representing 5–10% of the genome), but other "*Pst*I-like" clusters also exist (Burke and RajBhandari, 1982; Citterich et al., 1983; Burger and Werner, 1983). A comparison of the GC clusters from *N. crassa* and from *S. cerevisiae* suggests a similar evolutionary origin (i) in their amplification; (ii) in their construction; in the case of *N. crassa*, a series of duplication from an initial oligo(C) might have led to the formation of a "half-cluster" which could then be transformed into the present clusters by an inverted duplication; (iii) in their derivation from a primitive *ori* sequence postulated to share a basic oligo(C):oligo(G) element.

A situation apparently contradicting some of the points made above is that reported for *Torulopsis glabrata*, where no GC cluster was found in the short AT spacers forming the intergenic regions of the mitochondrial genome (Clark-Walker et al., 1985). It should be pointed out, however, that the primary structure of this genome is far from complete, and that a replication origin has not been localized. If the proposed identification of the *ori* sequence with a region characterized by several direct repeats is correct, a possibility worth considering is that this region is, in fact, a surrogate origin (see section **g**, below) replacing, with a lower efficiency, the original *ori* sequence.

(g) The function of GC clusters

As already mentioned, our original suggestion was that these newly found sequence elements were potential signals for the initiation of replication (Prunell and Bernardi, 1977); 24 GC clusters are, indeed, essential sequence elements of the *ori* sequences of the mitochondrial genome of yeast (de Zamaroczy et al., 1981) and four additional GC clusters (γ) are known to replicationally inactivate *ori* sequences 4, 6, 7 and 8; the role of two clusters β is not known, but their insertion seems to be required to preserve the tertiary structure of *ori* sequences which have undergone the insertion of cluster γ (de Zamaroczy et al., 1984). Other clusters, those of *var1, 15S, 21S* and, possibly, ORF5 play a coding role. This, however, leaves open the problem of the functional role of the other clusters, apart from their demonstrated use in recombination and petite genome excision (Fonty et al., 1978; de Zamaroczy et al., 1983). A number of suggestions have been made in this connection.

The similarity with the GC clusters of *ori* sequences, namely with signals which are definitely used in replication, suggested to us that at least some GC clusters could conserve in part the original functions. The finding of GC clusters of the *a1* family (namely of the family sharing the greatest similarity of secondary structure with the A-B region) in *ori*$^{\circ}$ petite genomes (which are deprived of canonical *ori* sequences), and the apparent correlation between their frequency in the repeat units and the suppressivity of the corresponding petites seemed to confirm that these sequences might play a role as replacement *ori* sequences (Goursot et al., 1982). Accordingly, it was proposed to call this family of clusters *ori*^{s}, for *ori* surrogate. A similar suggestion was made for some GC clusters resembling cluster C of *ori* sequences (Novitski et al., 1983). No direct evidence along this line is available yet and further experimentation is needed to reach a conclusion. In any case, petites deprived of any GC cluster can exhibit relatively high suppressivities (Fangman and Dujon, 1984; Foury, 1985) indicating that *ori*^{s} sequences are not needed for the replication of *ori*$^{\circ}$ petites.

Another role proposed for GC clusters (Tzagoloff et al., 1979; 1980; Novitski et al., 1983) has to do with the processing of primary transcripts. According to this hypothesis, the double-strand RNA structures corresponding to palindromic GC clusters or to palindromic sequences belonging in different GC clusters may provide cleavage sites for specific RNase implicated in RNA processing. Even if this hypothesis is attractive, the only evidence along this line is that two of the major cleavage points of the *oxi3-aap1-oli2*-ORF4 polycistronic transcript involve two clusters (Simon and Faye, 1984), namely 98 and 99. The similar case made for some *Pst*I site clusters on the *N. crassa* mitochondrial genome

(Citterich et al., 1983; Burger and Werner, 1983) may be accounted by the processing site being often associated with contiguous tRNA genes, as in the case of transcripts of mitochondrial DNAs from mammals (Anderson et al., 1981); and from *Schizosaccharomyces pombe* (Lang et al., 1983).

A different explanation, which we would like to propose here, is that the main or perhaps the only role for GC clusters is that of contributing to the structure and organization of the mitochondrial genome of yeast (and its AT spacers) and, indirectly, to the function of its regulatory elements. The suggestion is that the latter may be modulated by the genome configuration. As an example, one may quote the fact that certain promoter sequences are active whereas others, which do not differ from the former in primary structure, are not. For instance, (i) the nonanucleotide of sequence *r* is used to initiate the transcription-priming of *ori*2, but the identical nonanucleotide of sequence *r''* is not (Colin et al., 1985); (ii) sequence *i* is used to initiate the transcription of ORF5, but sequence *i'*, which is identical with sites used in the initiation of transcription of some tRNAs is not (Colin et al., 1985); (iii) the promoter sequence preceding the *cys* gene is identical with those used for the transcription of the *phe* and *fmet* genes and yet is inactive (Frontali et al., 1982; Christianson and Rabinowitz, 1983). This suggests that the signal sequences are necessary, but not sufficient elements for transcription and that the neighbouring sequences (and/or the secondary/tertiary structure of the region, and/or its "chromatin" structure) play a role in regulating their functional efficiency. This idea is basically similar to the construction of a three dimensional active site of an enzyme through the folding of the polypeptide chain and its higher-order structure. The argument has already been made (Bernardi, 1983) that the 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. The interstrain conservation of their primary structure (de Zamaroczy and Bernardi, 1985) provides an additional strong argument along the same line.

(h) Conclusions

The main points made in the present work are the following: (i) an analysis of the *ori* sequences has led us to propose a model for the evolutionary formation of *ori* sequences, in which a primitive *ori* sequence probably only made up by a monomeric cluster C and its flanking sequences, was the starting sequence for both a series of duplication-inversion events which generated clusters A and B and for a phenomenon of expansion which produced sequence *l*; incidentally, this model, inspired by the previous one proposed for the formation of intergenic sequences (Bernardi, 1982), suggests that the starting points for the formation of the latter are not the *ori* sequences but their evolutionary precursor(s); (ii) a comparison of GC clusters located outside and inside *ori* sequences indicates that the vast majority of the former (families *a1–a4*) are also originally derived from the primitive *ori* sequence in the course of its evolution towards the present *ori* sequences; in contrast, clusters from the other families seem to be derived from pre-existing clusters or very rarely, by statistical clustering of G's and C's (some clusters of the *n* group); (iii) finally, we have proposed that the function of GC clusters is predominantly, or entirely, associated with the structure and organization of the mitochondrial genome of yeast and, indirectly, with the regulation of its expression.

ACKNOWLEDGEMENTS

We thank Drs. R.P. Martin (Strasbourg) and R. Schweyen (Münich) for communicating to us their sequence results prior to publication. We thank Claude Mugnier for his help with computer programs, Martine Brient for typing this manuscript and Philippe Breton for the artwork.

REFERENCES

Anderson, S., Bankier, A.T., Barrell, B.G., De Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F.,Schreier, P.H., Smith, A.J.H., Staden, R. and Young, I.G.: Sequence and organization of the human mitochondrial genome. Nature 290 (1981) 457–465.

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.

Berlani, R.E., Bonitz, S.G., Coruzzi, G., Nobrega, M. and Tzagoloff, A.: Transfer RNA genes in the *cap-oxi1* region of yeast mitochondrial DNA. Nucl. Acids Res. 21 (1980) 5017–5030.

Bernardi, G.: Dimeric structure and allosteric properties of spleen acid deoxyribonuclease. J. Mol. Biol. 13 (1965) *603–605.*

Bernardi, G.: Mechanism of action and structure of acid deoxyribonuclease. Adv. Enzymol. 31 (1968) 1–49.

Bernardi, G.: Evolutionary origin and the biological function of noncoding 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., 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., Faures, 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., Piperno, G. and Fonty, G.: The mitochondrial genome of wild-type yeast cells. I. Preparation and heterogeneity of mitochondrial DNA. J. Mol. Biol. 65 (1972) 173–189.

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

Bordonné, R.: Structures primaires de gènes mitochondriaux de la levure *Saccharomyces cerevisiae*. Thèse de 3è cycle, Univ. L. Pasteur, Strasbourg, 1982.

Burger, G. and Werner, S.: Nucleotide sequence and transcript mapping of a mtDNA segment comprising *CO1*, tRNA-*arg* and several unidentified reading frames in *Neurospora crassa*, in Schweyen, R.J., Wolf, K. and Kaudewitz, F. (Eds.), Mitochondria – 1983: Nucleo-Mitochondrial Interactions. De Gruyter, Berlin, pp. 331–342.

Burke, J.M. and RajBhandary, U.L.: Intron within the large rRNA gene of *N. crasa* mitochondria: a long open reading frame and a consensus sequence possibly important in splicing. Cell 31 (1982) 509–520.

Butow, R.A.: Nonreciprocal exchanges in the yeast mitochondrial genome – A review. Trends Genet. 1 (1985) 81–84.

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.

Citterich, M.H., Morelli, G., Nelson, M.A. and Macino, G.: Expression of split genes of the *Neurospora crassa* mitochondrial genome, in Schweyen, R.J., Wolf, K. and Kaudewitz, F. (Eds.), Mitochondria – 1983; Nucleo-Mitochondrial Interactions. De Gruyter, Berlin, 1983, pp. 357–369.

Clark-Walker, G.D., Mc Arthur, 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.

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

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.

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., 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., 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.

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

Dujon, B.: Sequence of the intron and flanking exons of the mitochondrial *21S* rRNA gene of yeast strains having different alleles at the *ω* and *rib1* loci. Cell 20 (1980) 185–197.

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.

Fangman, W.L. and Dujon, B.: Yeast mitochondrial genomes consisting of only A-T base pairs replicate and exhibit suppressiveness. Proc. Natl. Acad. Sci. USA 81 (1984) 7156–7160.

Faugeron-Fonty, G., Le Van Kim, C., de Zamaroczy, M., Goursot, R. and Bernardi, G.: A comparative study of the *ori* sequences from the mitochondrial genomes of twenty wild-type yeast strains. Gene 32 (1984) 459–473.

Faugeron-Fonty, G., Mangin, M., Huyard, A. and Bernardi, G.: The mitochondrial genomes of spontaneous ori^r petite mutants of yeast have rearranged repeat units organized as inverted tandem dimers. Gene 24 (1983) 61–71.

Fonty, G., Crouse, E.J., Stutz, E. and Bernardi, G.: The mitochondrial genome of *Euglena gracilis*. Eur. J. Biochem. 54 (1975) 367–372.

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.

Foury, F.: A PIF-dependent recombinogenic signal in the mitochondrial DNA of yeast. EMBO J. (1985) submitted.

Frontali, L., Palleschi, C. and Francisci, S.: Transcripts of mitochondrial tRNA genes in *Saccharomyces cerevisiae*. Nucl. Acids Res. 10 (1982) 7283–7293.

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.

Goursot, R., Mangin, M. and Bernardi, G.: Surrogate origins of replication in the mitochondrial genomes of *ori*$^{\circ}$ 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.

Lang, B.F., Ahne, F., Distler, S., Trinkl, H., Kaudewitz, F. and Wolf, K.: Sequence of the mitochondrial DNA, arrangement of genes and processing of their transcripts in *Schizosaccharomyces pombe*, in Schweyen, R.J., Wolf, K. and Kaudewitz, F. (Eds.), Mitochondria – 1983: Nucleo-Mitochondrial Interactions, de Gruyter, Berlin, 1983, pp. 313–329.

Mangin, M., Faugeron-Fonty, G. and Bernardi, G.: The *ori*^{r} to *ori*$^{+}$ mutation in spontaneous yeast petites is accompanied by a drastic change in mitochondrial genome replication. Gene 24 (1983) 73–81.

Martin, R.P., Bordonné, R. and Dirheimer, G., The paromomycin in the yeast mitochondrial genome, in Akoyounoglou, G., Evangelopoulos, E.A., Georgatsos, J., Palaiologos, G., Trakatellis, A. and Tsiganos, C.P. (Eds.), Cell Function and Differentiation, Part B. Liss, New York, 1983, pp. 355–365.

Mery-Drugeon, E., Crouse, E.J., Schmitt, J.M., Bohnert, H.-J. and Bernardi, G.: The mitochondrial genomes of *Ustilago cynodontis* and *Acanthamoeba castellanii*. Eur. J. Biochem. 114 (1981) 577-583.

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

Miller, D.L., Underbrinck-Lyon, K., Najarian, D.R., Krupp, J. and Martin, N.C.: Transcription of yeast mitochondrial tRNA genes and processing of tRNA gene transcripts, in Schweyen, R.J., Wolf, K. and Kaudewitz, F. (Eds.), Mitochondria – 1983: Nucleo-Mitochondrial Interactions, de Gruyter, Berlin, 1983, pp. 151–164.

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.

Novitski, C.E., Macreadie, I.G., Maxwell, R.J., Lukins, H.B., Linnane, A.W. and Nagley, P.: Features of nucleotide sequences in the region of the *oli2* and *aap1* genes in the yeast mitochondrial genome, in Nagley, P., Linnane, A.W., Peacock, W.J. and Pateman, J.A. (Eds.), Manipulation and Expression of Genes in Eukaryotes, Academic Press, Sydney, 1983, pp. 257–268.

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., 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.

Schmitt, J.M., Bohnert, H.-J., Gordon, K.H.J., Herrmann, R., Bernardi, G. and Crouse, E.J.: Compositional heterogeneity of the chloroplast DNAs from *Euglena gracilis* and *Spinacia oleracea*. Eur. J. Biochem. 117 (1981) 375–382.

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.

Simon, M. and Faye, G.: Organization and processing of the mitochondrial *oxi3/oli2* multigenic transcript in yeast. Mol. Gen. Genet. 196 (1984) 266–274.

Sor, F. and Fukuhara, H.: Nature of an inserted sequence in the mitochondrial gene coding for the *15S* ribosomal RNA of yeast. Nucl. Acids Res. 11 (1982) 1625–1633.

Tzagoloff, A., Macino, G., Nobrega, M.P. and Li, M.: Organization of mitochondrial DNA in yeast, in Cummings, D.J., Borst, P., David, I.B., Weissman, S.M. and Fox, C.F. (Eds.), Extrachromosomal DNA, Academic Press, New York, 1979, pp. 339–355.

Tzagoloff, A., Nobrega, M., Akai, A. and Macino, G.: Assembly of the mitochondrial membrane system. Organization of yeast mitochondrial DNA in the *oli1* region. Curr. Genet. 2 (1980) 149–157.

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.

Wong, J.F.H., Ma, D.P., Wilson, R.K. and Roe, B.A.: DNA sequence of the *Xenopus laevis* mitochondrial heavy and light strand replication origins and flanking tRNA genes. Nucl. Acids Res. 11 (1983) 4977–4995.

Yin, S., Heckman, J. and RajBhandary, U.L.: Highly conserved GC-rich palindromic DNA sequences flank tRNA genes in *Neurospora crassa* mitochondria. Cell 26 (1981) 325–332.

Communicated by M.R. Culbertson.

NOTE ADDED IN PROOF

Recent mapping data have shown that cluster 120 is in fact nothing but cluster 132 as read at the opposite end of an independent petite genome.