Achievements and Perspectives of Mitochondrial Research
Volume II: Biogenesis, E. Quagliariello et al. editors

THE ORGANIZATION AND EVOLUTION OF THE MITOCHONDRIAL GENOME OF YEAST : ORI SEQUENCES, GENES AND INTERGENIC SEQUENCES

MIKLOS DE ZAMAROCZY AND GIORGIO BERNARDI
Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2, Place Jussieu, 75005 Paris, France.

INTRODUCTION

Although the goal of establishing the complete primary structure of the mitochondrial genome from wild-type Saccharomyces cerevisiae has not yet been reached, 90% of the sequence is now known (all values quoted in this paper are in round figures). We report here an analysis of the available data (1), with a special emphasis on the 190 GC clusters present in the sequences (M. de Zamaroczy and G. Bernardi, paper submitted for publication) and discuss the results in connection with the problem of the organization and evolution of the mitochondrial genome of yeast.

THE SEQUENCED REGIONS, THE SIZE AND NUMBER OF THE MITOCHONDRIAL GENOME UNITS

A critical compilation of the regions sequenced so far in a number of S.cerevisiae strains indicates that they comprise 74,000 bp (base pairs) in long genomes (which contain a full complement of intervening sequences), and 69,000 bp in short genomes. Sequence gaps (Fig. 1) belong to two classes : (i) intra-regional gaps are gaps within sequenced regions; they amount to a total of 3,500 bp for long genomes, and 1,500 bp for short genomes; (ii) inter-regional gaps are gaps which exist between independently sequenced regions; they amount to 8,000 bp in both long and short genomes.

The sum of sequenced regions and gaps leads to an estimate of 85,000 bp for the size of long genomes; by subtracting introns aI5α and β and bI1-3, one can estimate the size of short genomes as 78,500 bp; by further subtracting introns aI1, aI4 and rI one obtains a size of 74,000 bp for the supershort genome of S.carlsbergensis. All these estimates are 7-11% higher than the previous ones, which were based on restriction maps.

Using the relative amount of mitochondrial DNA determined for strain A (2), 13.3%, and a nuclear genome size of 9 ± 2.10^9 (3), the

copy number of long mitochondrial genome units of haploid yeast can be estimated as 21-33, a value about half that usually quoted of 50.

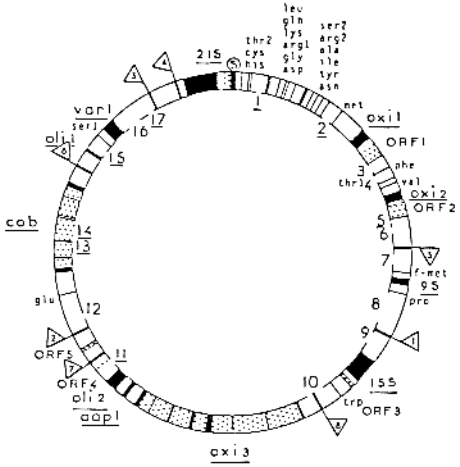

Fig. 1. Physical map of a long mitochondrial genome unit of wild-type S.cerevisiae showing the remaining sequence gaps (1); these are numbered 1 through 17; underlined numbers refer to intra-regional gaps, other numbers to inter-regional gaps (see Text). Numbers in triangles indicate the location of 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 (8). Ⓢ corresponds to the SalI site used as the origin of the map.

INTERSTRAIN DIFFERENCES IN PRIMARY STRUCTURE

Apart from the differences due to facultative introns, to facultative mini-inserts in the 15S, 21S and var1 genes (these correspond to GC clusters and, in the latter case, also to two short (AAT)n sequences), and point mutations in coding sequences, interstrain differences concern intergenic sequences. Large changes have been found in three regions : ori1-15S (400 bp), oli2-ORF4 (1800 bp) and ori4 (in this case, 1900 bp are missing in just one strain out of 20). Otherwise, changes essentially consist of small insertions/deletions and rearrangements. Remarkably, however, if these changes are neglected, the divergence in the 7-8% of intergenic sequences determined so far in different strains is only 1-2%. A high level of interstrain homology in intergenic sequences had been previously indicated by the very similar numbers of HaeIII and HpaII sites, which essentially correspond to certain families of GC clusters and by the similar genomic distribution of these sites (4). These lines of evidence justify putting together sequence results obtained from different strains, as we have done (1).

CODING AND NON-CODING SEQUENCES

The data of Table I indicate that an actual coding role has been demonstrated so far for only 16% of the long genomes. This value can be increased to about 30% if one takes into account all intronic ORF's, and to about 34% by considering, in addition, the intergenic ORF's. In other words, even by maximally stretching the potential coding sequences, no more than 1/3 of the mitochondrial genome of yeast is made up of coding sequences. The case for a coding role of intergenic ORF's, except for ORF5, is, however, very weak (5); moreover, the genetic evidence for a coding role of intronic ORF's is only available for 80% of their overall amount. This indicates that coding sequences might well be closer to 1/4 than to 1/3 of the mitochondrial genome.

TABLE I

Size of mitochondrial DNA sequences

	Size %
1. Total genome	100
2. Genes or exons	16.2
3. Introns	21.6
a. ORF	15.3
b. CRF	6.3
4. Intergenic regions	62.2
a. Intergenic ORF	3.9
b. *Ori* sequences	2.6
Coding regions (maximal)	35.4
Noncoding regions (minimal)	64.6

THE GC CLUSTERS

The number of GC clusters (4,6) in the sequenced regions of the mitochondrial genome of yeast is 190; GC clusters represent 9% of the sequences. On the basis of their primary structures GC clusters can be divided into eight classes. The first class is formed by clusters A, B and C of *ori* sequences; as shown in Fig. 2, these clusters are related to each other (see also below). The vast majority of the other clusters belong to four closely related families, *a1*-*a4*; most of these contain sequence elements quasi-identical with clusters A and B of *ori* sequences (Fig. 3) and can be folded into stem-and-loop structures. The other families are the *v* family, which is related to the *a2* family and can be folded into a stem-and-loop structure; the *c* family, which is related to the C clusters of *ori* sequences; and the *n* group, which is formed by short clusters which cannot be directly correlated with the other families.

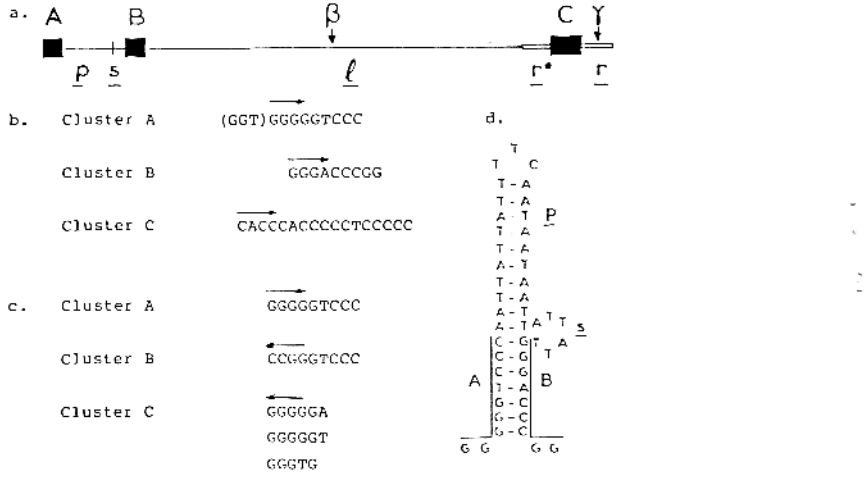

Fig. 2. (a) A schematic representation of an ori sequence (8). The position of clusters β and γ, which are only present in some ori sequences, are indicated. (b) The GC clusters A, B and C from the ori sequences of the mitochondrial genome of yeast. The primary structure is read on the strand carrying the oligo-C stretches of cluster C. (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.

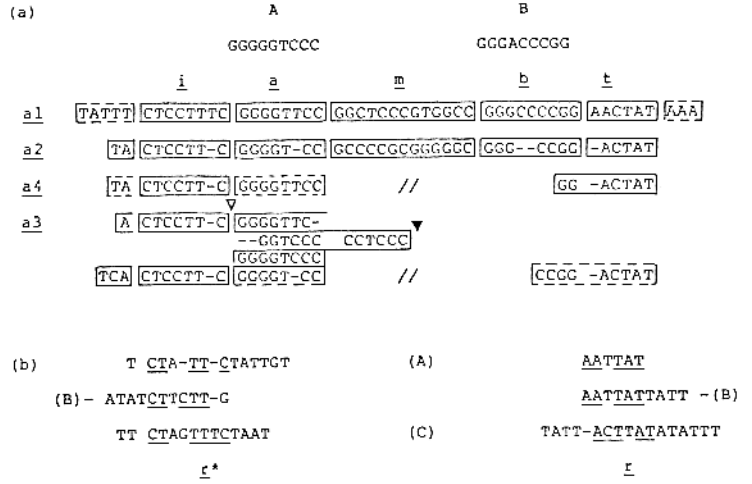

Fig. 3. (a) A comparison of the primary structures of GC clusters from the a1-a4 families. Common sequence elements are boxed (with broken lines when sequences show more variability) and compared with GC clusters A and B from ori sequences. Double dashes stand for variable sequences; open and closed triangles indicate two short inserts. (b) Sequences flanking cluster 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 ori1; a lower homology is found for ori2 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 ori2, the other ori sequences showing lower homologies. The t homologous sequences are identical in all ori sequences, except for the insertion of clusters γ in r sequence of ori4, 6-8, and for the fact that the sequence precedes instead of following cluster B.

GC CLUSTERS A, B, C AND THE EVOLUTIONARY ORIGIN OF ORI SEQUENCES

The most ancient sequence element of ori sequences is likely to be cluster C (or its monomeric penta C) and its contiguous sequences r* and r, (i) since bidirectional replication is initiated at r* and r by RNA primers and continued into nascent DNA chains at the level of cluster C (7) and (ii) since the latter is an evolutionarily conserved sequences identified (8) with sequence CSB-2 of vertebrate mitochondrial genomes. The formation of an ori sequence might then have occurred (Fig. 4) through (i) a duplication of the postulated primitive ori sequence, formed by the r*-penta C-r system, followed by the functional inactivation of the copy; (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; (iv) the formation of cluster A by an inverted duplication of the penta C copy; (v) the formation of the A-B region through the duplication-inversion of cluster A and its flanking AT region.

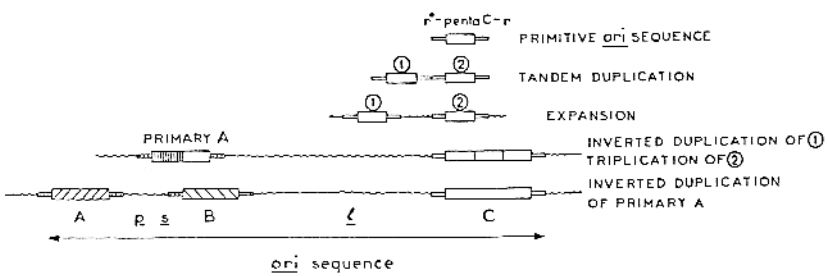

Fig. 4. Hypothetical scheme for the evolutionary construction of ori sequences. See text for details.

THE EVOLUTIONARY ORIGIN OF THE OTHER GC CLUSTERS

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. The primary clusters so originated might then have undergone amplification (by duplication and translocation); the high sequence homology of cluster families is suggestive of amplification bursts; the non-randomness of flanking sequences (Fig. 3) of a targeting towards acceptor sequences.

In contrast, clusters from the other families seem to be derived from pre-existing clusters of the a2 family (case of v clusters), from cluster C (case of c clusters), from pre-existing clusters of all families by recombination events (case of the a4 family and n group), or, possibly, by statistical distribution of G's and C's (some clusters of the n group).

GC clusters located outside ori sequences are overwhelmingly present in intergenic sequences. The exceptions are six clusters in intergenic ORF1-4, four clusters in mitochondrial genes (15S, 21S, var1 and ORF5), 16 clusters in intronic CRF's and two clusters in two intronic ORF's. The former probably should be considered just as other intergenic clusters for the reasons given above. The GC clusters of the 15S and 21S genes are facultative inserts found in some strains only; in contrast, those of var1 and ORF5 reflect a different situation. Indeed, when the 1 kb sequence of var1 was compared with 1 kb of intergenic sequences around ori1, it was found that these stretches have comparable high numbers of direct and inverted repeats and many common sequences (9). This finding favors the idea that var1 is a gene which arose recently from an intergenic sequence (10), a view also supported by the fact that var1 is not even conserved in another ascomycete, Neurospora crassa. The evolutionary story of ORF5 might well be the same. As far as the intronic clusters are concerned, they will be discussed elsewhere.

THE FUNCTION OF GC CLUSTERS

Our initial suggestion was that GC clusters were signals for the initiation of replication and transcription (6). While the latter suggestion was based on hypotheses which were later disproven, 24 GC clusters are, indeed, present in the eight ori sequences (8,11); four additional clusters, γ, are known to replicationally inactivate ori4,6-8 and the two clusters β seem to be required

to preserve the tertiary structure of ori sequences carrying cluster γ (8). The clusters located in genic sequences play a coding role. As far as all other clusters are concerned, their function is unknown, apart from their demonstrated role in recombination and petite genome excision (12,13). A number of suggestions have been made in this connection.

The similarity with the GC clusters of ori sequences suggested to us that at least some GC clusters could conserve in part the original functions. The finding of GC clusters of the al family (namely of the family sharing the greatest similarity of secondary structure with the A-B region) in ori° 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 (14). 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 (15). No direct evidence along this line is available yet, however. In any case, petites deprived of any GC cluster can exhibit relatively high suppressivities (16, F. Foury, personal comm.) indicating that ori^{S} sequences are not needed for the replication of ori° petites.

Another role proposed for GC clusters (15,17,18) has to do with the processing of primary transcripts. According to this hypothesis, the double-stranded RNA regions corresponding to palindromic GC clusters or to palindromic sequences belonging in different GC clusters may provide cleavage sites for specific RNases implicated in RNA processing. Even if this hypothesis is attractive, no evidence along this line has been reported either.

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, indirectly, to the function of its regulatory elements. The suggestion is that the latter may be modulated by the genome configuration. In favor of this idea, one may quote the fact that certain promotor sequences are active whereas other ones, 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 ori2, but the identical nonanucleotide of sequence r" is not (5); (ii) a sequence (called i, 5) is used to initiate the transcription of ORF5, but another one (called i'), which is identical with sites used in the initiation of transcription of some tRNAs is not; (iii) the promotor sequence preceding the cys gene is identical with those used for the transcription of the phe and fmet genes and yet is inactive (19,20). This suggests that the signal sequence itself is a necessary, but not a sufficient element for transcription and that the neighboring sequences, and/or the secondary/tertiary structure of the region, and/or its "chromatin" structure also play a role. This idea is basically similar to the construction of a tridimensional active site of an enzyme through the folding of the polypeptide chain and its higher order structure. The argument has already been made (10,21) that the conservation in amount (and in primary structure; 1) 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.

REFERENCES

1. de Zamaroczy M, Bernardi G (1985) Gene 37 : 1-17.
2. Bernardi G, Piperno G, Fonty G (1972) J Mol Biol 65 : 173-189.
3. Lauer GD, Roberts TM, Klotz LC (1977) J Mol Biol 114 : 507-526.
4. Prunell A, Kopecka H, Strauss F, Bernardi G (1977) J Mol Biol 110 : 17-52.
5. Colin Y, Baldacci G, Bernardi G (1985) Gene 36 : 1-13.
6. Prunell A, Bernardi G (1974) J Mol Biol 86 : 825-841.
7. Baldacci G, Chérif-Zahar B, Bernardi G (1984) EMBO J 3 : 2115-2120.
8. de Zamaroczy M, Faugeron-Fonty G, Baldacci G, Goursot R, Bernardi G (1984) Gene 32 : 439-457.
9. Bernardi G, Bernardi G (1980) FEBS Lett 115 : 159-162.
10. Bernardi G (1983) Folia Biol 29 : 82-92.
11. Faugeron-Fonty G, Le Van Kim C, de Zamaroczy M, Goursot R, Bernardi G (1984) Gene 32 : 459-473.
12. Fonty G, Goursot R, Wilkie D, Bernardi G (1978) J Mol Biol 119 : 213-235.
13. de Zamaroczy M, Faugeron-Fonty G, Bernardi G (1983) Gene 21 : 193-202.
14. Goursot R, Mangin M, Bernardi G (1982) EMBO J 1 : 705-711.
15. Novitski CE, Macreadie IG, Maxwell RJ, Lukins HB, Linnane AW, Nagley P (1984) Curr Genet 8 : 135-146.
16. Fangman WL, Dujon B (1984) Proc Acad Sci USA 81 : 7156-7160.

17. Tzagoloff A, Macino G, Nobrega MP, Li M (1979) In : Cumings D et al. (eds) Extrachromosomal DNA. Academic Press, New York, pp 339-355.

18. Tzagoloff A, Nobrega MP, Akai A, Macino G (1980) Curr Genet 2 : 149-157.

19. Frontali L, Palleschi C, Francisci S (1982) Nucl Acids Res 10 : 7283-7293.

20. Christianson T, Rabinowitz M (1983) J Biol Chem 258 : 14025-14033.

21. Bernardi G (1982) In : Slonimski PP, Borst P, Attardi G (eds) Mitochondrial Genes. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp 269-278.