Reprinted from *Mitochondrial Genes*

Evolutionary Origin and the Biological Function of Noncoding Sequences in the Mitochondrial Genome of Yeast

Giorgio Bernardi
Laboratoire de Génétique Moléculaire
Institut de Recherche en Biologie Moléculaire
75005 Paris, France

During the past 15 years we have carried out in our laboratory a series of investigations on the mitochondrial genome of *Saccharomyces cerevisiae*. Very early, the molecular approaches used led us to the discovery that the majority of this genome is formed by noncoding sequences. We have since concentrated our efforts on the organization of such sequences and on their biological role. In this paper I summarize what we have learned from this work and then present new results that shed some light on the evolutionary origin and the function of these sequences.

ORGANIZATION OF THE MITOCHONDRIAL GENOME OF YEAST

Sequence work performed in several laboratories during the past 3 years has confirmed our previous conclusions (for a brief review, see Bernardi 1979) on the organization of the mitochondrial genome of yeast. The majority of this genome is formed by (1) long AT stretches made up of short dAT:dAT and dA:dT sequences with rare dG:dC base pairs; AT stretches are internally repetitive in sequence and rich in palindromes (Bernardi and Bernardi 1980); and (2) short GC clusters characterized by sequences that often are symmetric and largely homologous to each other; GC clusters are embedded in AT stretches (Cosson and Tzagoloff 1979; Gaillard and Bernardi 1979). As predicted, repeated sequences within AT stretches and GC clusters are used as sites for the excision of the defective genomes of spontaneous petite mutants (Baldacci et al. 1980; Gaillard et al. 1980). Expectedly, AT stretches and GC clusters form the intergenic regions of the mitochondrial genome (Fig. 1); in addition, they are also present in the intervening sequences of the *cob* and *oxi3* genes (discussed later).

A computer analysis extending an original work (Bernardi and Bernardi 1980) to over 20,000 bp from AT stretches and GC clusters (see Evolutionary Origin of Intergenic Sequences) has confirmed our previous conclusion that these sequences share common features in that they are basically made up of the same repeated sequences. Two major questions

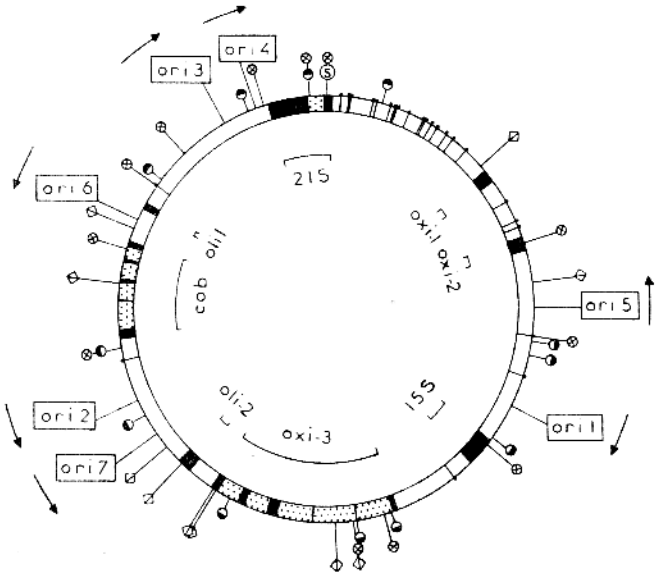

Figure 1 Physical map of the mitochondrial genome of the wild-type strain A used in the present work (see Faugeron-Fonty et al. 1979). (■) Coding sequences for the genes indicated on the inner circle; (▨) intervening sequences. tRNA genes are indicated by thin radial lines. The localization of *ori* sequences is given together with their orientation (arrows point in the direction of cluster C). Restriction sites: (Ⓢ) *Sal*I; (Ⓧ) *Hin*cII; (◁▷) *Eco*RI; (◓) *Hha*I. (Reprinted, with permission, from de Zamaroczy et al. 1981.)

are then raised: (1) How could a common sequence pattern arise at many different locations in the genome? (2) What biological role(s), if any, can be fulfilled by AT stretches and GC clusters?

The first problem, concerning the evolutionary origin of the intergenic sequences (see below), could be approached after we learned about the origins of DNA replication in the mitochondrial genome of yeast and about their own evolutionary origin (see below). As for the second problem, we already knew about the role of AT stretches and GC clusters in the excision of the genomes of spontaneous petite mutants (Faugeron-Fonty et al. 1979) and in site-specific recombination events, which are so frequent in yeast crosses (Fonty et al. 1978) and which also seem to occur in vegetative cells, giving rise to the polymorphism of mitochondrial genomes from different strains (Prunell et al. 1977). In addition, we suggested that sequences from AT stretches and GC clusters, whose symmetry features are indicative of specific protein-binding ability, might play regulatory roles associated with DNA replication and transcription (Prunell and Bernardi 1977). This second point is discussed later (see Biological Role of Intergenic Sequences).

ori SEQUENCES AND THEIR EVOLUTIONARY ORIGIN

We had previously suggested (Prunell and Bernardi 1977; Bernardi et al. 1980) that the mitochondrial genome of yeast contains several origins of replication, at least one of which is present in the excised genomic segment that becomes the repeat unit in the genomes of spontaneous petites. Therefore, when we first sequenced two such genomes derived from the same region of the wild-type genome, we looked for an origin of DNA replication within their repeat units and focused our attention on a common central region characterized by three GC clusters. Two of these clusters, A and B, could be folded into a hairpin loop, and the third cluster, C, was made up of a polypyrimidine:polypurine stretch separated from the other two clusters by a long AT region (Fig. 2) (Gaillard and Bernardi 1979; Bernardi et al. 1980; de Zamaroczy et al. 1980, 1981). The folding of clusters A and B and the sequence of cluster C were very similar to structures found in the origins of replication of mammalian mitochondrial genomes (Crews et al. 1979; Gillum and Clayton 1979; Kobayashi et al. 1980). An *ori* sequence, as we called it, was found in the vast majority of the genomes of spontaneous petites. Partial or total deletions of such an *ori* sequence, or its rearrangement, profoundly affected the suppressivity of the corresponding petite, which we showed to correspond to the level of transmission of the petite genome to the progeny of its cross with wild-type cells (Bernardi et al. 1980; Goursot et al. 1980; de Zamaroczy et al. 1981).

Mapping and hybridization studies (de Zamaroczy et al. 1981) provided evidence for at least seven *ori* sequences in the mitochondrial genomes of wild-type cells. All *ori* sequences are extremely homologous to each other, except that some of them contain additional GC clusters identical in sequence and localization (Fig. 2). Recent investigations (G. Baldacci and G. Bernardi, unpubl.) have shown that *ori* sequences are also used as origins of transcription in petite genomes; such transcription is lost in genomes whose *ori* sequences are fully deleted or in genomes that lack cluster C. In this connection, it is of interest that different *ori* sequences have different orientations in the wild-type genome, since this suggests that both strands of the latter are transcribed.

As far as the evolutionary origin of *ori* sequences is concerned, it is difficult to escape the conclusion that they arose as the result of duplication and translocation events. Selective advantages provided by these events might be faster replication (probably related to the need to rapidly increase the copy number of the few mitochondrial genome units segregated into the buds) and a finer regulation of gene expression (see Biological Role of Intergenic Sequences).

EVOLUTIONARY ORIGIN OF INTERGENIC SEQUENCES

If one considers that five of the eight intergenic regions of the yeast mitochondrial genome contain one or two *ori* sequences (Fig. 1; tRNA

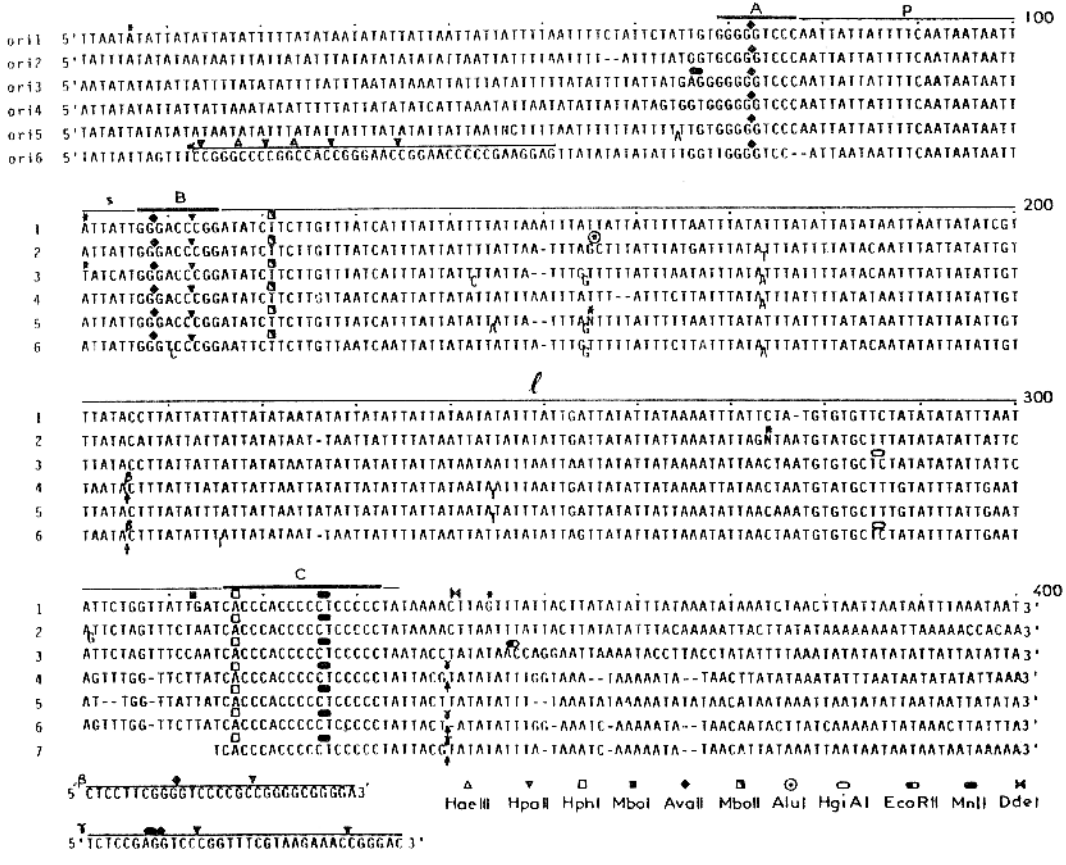

Figure 2 Primary structure of the *ori* sequences and their flanking regions. (—) GC clusters; (—) AT sequences within the *ori* sequences. (Reprinted, with permission, from de Zamaroczy et al. 1981; see this reference for additional information.)

genes are disregarded); that one or another *ori* sequence may be absent from the genome of a given strain (G. Faugeron-Fonty, pers. comm.), a fact possibly accounting for the absence of an *ori* sequence in some intergenic regions; that the presence of one (or more) additional *ori* sequences is not ruled out; and that *ori* sequences and intergenic sequences are both made up of AT spacers and GC clusters, then a reasonable working hypothesis is that intergenic sequences derived from *ori* sequences by an expansion phenomenon. Such an expansion may have taken place through two different mechanisms, which may represent the first and second steps in the process: (1) a slippage of the replicase at the *ori* sequence; this is a well-known phenomenon for the reiterative replication of poly(dAT:dAT) by DNA polymerase I (Kornberg et al. 1964); and (2) unequal crossovers; evidence for the high frequency of such a phenomenon in mitochondrial recombination is available (Fonty et al. 1978).

Such a working hypothesis was tested by comparing the *ori* sequences with noncoding sequences using a computer program set up by J. Ninio (Institut de Recherche en Biologie Moléculaire, Paris, France). These noncoding sequences, external to *ori* sequences, comprised two sets (A and B) of about 10,000 nucleotides each, derived from the 15S–*oli2* and the *oli2*–15S segments of the mitochondrial genome (see Fig. 1), respectively; their primary structure has been described by Tzagoloff and his colleagues. Such a comparison showed, for instance, that 10% of the noncoding sequences of set A were formed by sequences 12 nucleotides long (or longer), which were present in *ori3*. In contrast, only 1% of the same noncoding sequences were formed by sequences 12 nucleotides long (or longer), which were present in random sequences having the length and the base composition of *ori3*. As expected, similar results were obtained when testing either set of noncoding sequences against any *ori* sequence. Since different *ori* sequences are not identical with each other, it is useful to compare external noncoding sequences not with single *ori* sequences, but with several or all of them. When three *ori* sequences were considered, 20% of the noncoding sequences (set A) were found to be formed by sequences as long as, or longer than, 12 nucleotides present in the *ori* regions. Preliminary data suggest that if all *ori* sequences are tested, about 30% of each one of the two sets of reference base pairs will be present in such sequences, the values for control random sequences still being close to 10% of those of *ori* sequences. If one looks at the distribution of such repeats in the *ori* sequences, one can see that they derive from all their sections (Fig. 3). GC clusters, however, seem to be underrepresented. An obvious reason for this is the fact that GC clusters are relatively short; they were tested at the level of sequences equal to, or longer than, 7 nucleotides. When this is done, sequences from all GC clusters of *ori* sequences are also found in large number outside the *ori*

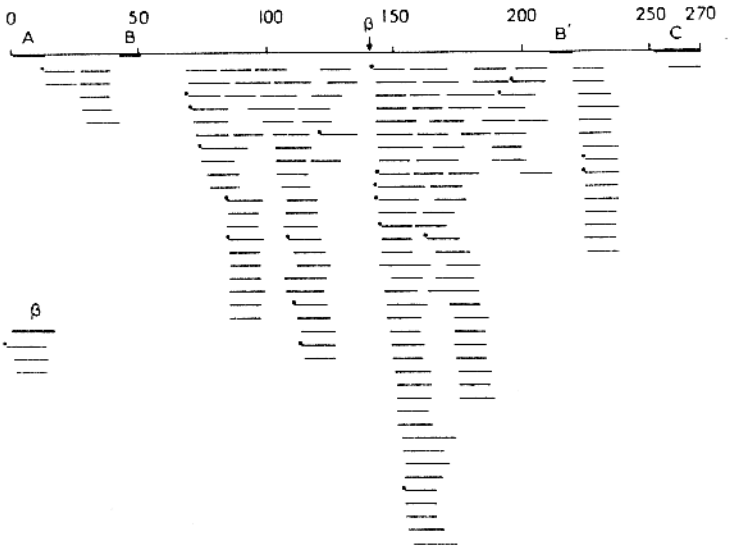

Figure 3 Sequences (≥12 nucleotides) from three *ori* sequences found in intergenic and intervening (●) sequences from the mitochondrial genome of yeast (A set; see text). Cluster B corresponds to a region relatively rich in GC, present at the location indicated by the positions on the abscissa.

sequences (Fig. 4). We conclude, therefore, that the idea that intergenic sequences are the result of an expansion of *ori* sequences is a plausible one.

BIOLOGICAL ROLE OF INTERGENIC SEQUENCES

Before considering this issue, it should be noted that the mitochondrial genome of yeast shares with the nuclear genomes of eukaryotes two fundamental properties: (1) the c-value paradox, i.e., the excess of DNA over the amount required for coding gene products; and (2) the sequence organization, which is characterized by an interspersion of unique or single-copy sequences with repeated and palindromic sequences. Thus, any conclusion about the biological role of repeated and palindromic sequences in the mitochondrial genome of yeast may also apply to the general case of the nuclear genomes of eukaryotes. As far as the biological role of the intergenic sequences in the mitochondrial genome of yeast is concerned, two extreme possibilities should be considered. The first is that such sequences are "selfish" or "junk" DNA (Orgel and Crick 1980; Doolittle and Sapienza 1980). This possibility is of special relevance because the mechanism by which repeated and foldback sequences origi-

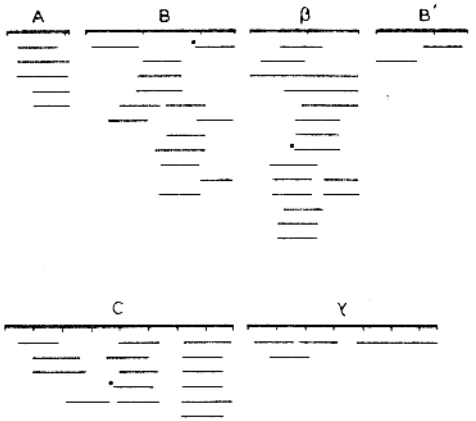

Figure 4 Sequences (≥7 nucleotides) from the GC clusters of *ori6* found in intergenic and intervening (●) sequences from the mitochondrial genome of yeast (A set; see text).

nated in the mitochondrial genome appears to be the one postulated for the origin of selfish DNA, i.e., an expansion of replication origins (Orgel and Crick 1980). The second possibility is that those sequences fulfill a physiological role. It is worth stressing here that (1) the two possibilities just mentioned are not mutually exclusive, since each one of them might account for a fraction of the intergenic sequence; and (2) much more precise statements can be made about a number of questions in the case of the mitochondrial genome of yeast than in the case of the nuclear genome of eukaryotes.

In the first case, there are several arguments against the idea that intergenic sequences correspond, at least to a large extent, to selfish DNA. First of all, the increase in the number of *ori* sequences and, more so, their expansion lead to increased levels of sequence homology in the genome, causing a very considerable instability. Such a fact is extremely well documented, with direct repeats within *ori* sequences or in intergenic sequences being used as excision sites for the formation of petite genomes (Baldacci et al. 1980; Gaillard et al. 1980). It is obvious that such a price—genomic instability—can only be paid if there is a compensatory selective advantage. A similar argument could be applied to the increased energy expenditure and longer time associated with the replication of additional DNA (the longer genomic units, as soon as they arise, would be competed out by the shorter ones, as shown by similar competition events found with petite genomes [de Zamaroczy et al. 1981]). Second, the expansion of *ori* sequences does not propagate nonsense sequences;

rather, it propagates sequences that have been highly selected and conserved in evolution and whose primary role is to interact specifically with proteins involved in DNA replication and transcription. In other words, the expansion of *ori* sequences leads to the propagation of potential regulatory signals, which will be used whenever they provide a selective advantage. As an example of the use of such sequences, *ori*^{0} petite genomes (which lack an *ori* sequence) contain GC clusters homologous to those of *ori* sequences; such clusters are likely to act as surrogate *ori* sequences. On the other hand, repression and derepression phenomena and nucleo-mitochondrial interactions require a number of signals involved in the regulation of gene expression. Similarly, the complex maturation pattern of primary transcripts in the mitochondrial genome (Van Ommen et al. 1979) requires a large number of precise processing signals; the role of GC clusters in this connection has been mentioned in view of their specific sequence features (Tzagoloff et al. 1979). Third, repeated and palindromic sequences in intergenic regions play a role in mitochondrial recombination and also, in all likelihood, in excision, insertion, and translocation events which provide evolutionary adaptability.

Incidentally, all of the arguments presented above support the idea that the mitochondrial genome of yeast is more "evolved" than the genomes of animal cells. This can be easily understood if one considers the much higher physiological flexibility required of a free-living cell able to survive under a number of conditions, including anaerobiosis, compared with that required of an animal cell.

EVOLUTIONARY ORIGIN AND THE BIOLOGICAL ROLE OF INTERVENING SEQUENCES

Examination of the existing sequence data (Fig. 5) shows that the intervening sequences of the mitochondrial genes *cob* and *oxi3*, coding for cytochrome *b* and subunit I of cytochrome oxidase, have a tripartite structure with a middle closed reading frame separating two open reading frames contiguous to the sequences coding for the subunits of the two respiratory proteins. Some of the open reading frames in intervening sequences appear to code for proteins involved in the processing of primary transcripts (Lazowska et al. 1980). The closed reading frames show the same sequence features found in intergenic regions. A computer analysis of the sequences of these closed reading frames (Figs. 3 and 4) shows that sequences from within the *ori* sequences are present at exactly the same extent as in intergenic sequences. Nothing differentiates the closed reading frames of intervening sequences from the intergenic sequences. This may mean that the closed reading frames of intervening sequences are evolutionary remnants of intergenic sequences separating different ancestral genes which have become transcriptionally linked with each other.

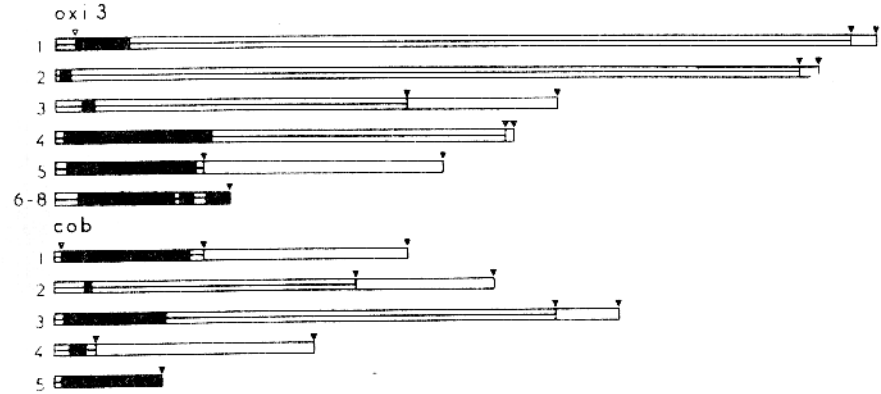

Figure 5 Sequence organization of the *oxi3* and *cob* genes. (■) Sequences used for coding cytochrome oxidase subunit I and cytochrome *b*, respectively; (▭) open reading frames; (□) closed reading frames; (▽) codons for methionine; (▼) ocher codons delimiting closed reading frames.

Obviously, the *ori* sequences which originated the present intervening sequences have been lost. As far as the biological role of the closed reading frames is concerned, recent observations (C. Jacq et al., pers. comm.) have shown that mutations in them affect the excision of the intervening sequences. This leads to the same overall conclusion already reached for intergenic sequences, i.e., that they also play a regulatory role.

In conclusion, the evidence available at the present time appears to support the idea that the complex sequence organization of the mitochondrial genome of yeast corresponds to the needs of very active and finely regulated replication, transcription, and recombination processes. This seems to be achieved at the price of exceptional genomic instability.

ACKNOWLEDGMENTS

Thanks are due to Gregorio Bernardi, Jean-Pierre Dumas, and Jacques Ninio for their generous help with the computer analysis reported here; to Philippe Breton for the art work; and to Martine Brient for typing the manuscript.

REFERENCES

Baldacci, G., M. de Zamaroczy, and G. Bernardi. 1980. Excision sites in the GC clusters of the mitochondrial genome of yeast. *FEBS Lett.* **114:** 234.

Bernardi, G. 1979. The petite mutation in yeast. *Trends Biochem. Sci.* **4:** 197.

Bernardi, G. and G. Bernardi. 1980. Repeated sequences in the mitochondrial genome of yeast. *FEBS Lett.* **115:** 159.