REPEATED SEQUENCES IN THE MITOCHONDRIAL GENOME OF YEAST

Gregorio BERNARDI and Giorgio BERNARDI

Ecole Nationale Supérieure des Télécommunications, 46, Rue Barrault, 75013 Paris and *Laboratoire de Génétique Moléculaire, Institut de Recherche en Biologie Moléculaire, 2 Place Jussieu, 75005 Paris, France*

Received 6 May 1980

1. Introduction

The existence of stretches rich in dAT:dAT and dA:dT in the mitochondrial genome of *Saccharomyces cerevisiae* was first reported in 1968 for the genome of a 'petite' mutant [1]. Investigations carried out in the following years (reviewed [2]) not only characterized the AT-rich stretches (or AT spacers) of the mitochondrial genome of wild-type yeast cells in their amount ($\geqslant$50% of the genome), GC level ($<$5% GC), and pyrimidine isostichs, but also suggested that sequence repetitions in AT spacers (and in the GC clusters later found to be embedded in them) could account for the great instability of the genome. Recent work [2–6] has demonstrated that direct repeats 13–23 nucleotides long in the AT spacers and in the GC clusters are indeed used in illegitimate site-specific recombination events to excise mitochondrial DNA segments, which then become the repeat units of 'petite' mutants.

The availability of the primary structure of two long AT-rich segments allowed us to study the sequence repetitions and the palindromes in those stretches. Segment I (fig.1) is the repeat unit of the mitochondrial genome of 'petite' mutant a-1/1R/1 [3,4]; this originated from the region around the 15 S RNA gene [3,7]. Segment II (fig.1) is *Hpa* II fragment 2 of 'petite' mutant DS 401 [8]; this was the putative locus of the *var 1* gene and is located roughly opposite to segment I on the circular map of the mitochondrial genome.

2. Methods

In order to gather information about repeated sequences, a Fortran program was written for a Honeywell-Bull Iris 80 computer. Basically, this first searched for all repeated tetranucleotides, and then used this set of data to find out longer repeats by a recurrence method. Inverted repeats, palindromes and repeats with mismatch were searched out on the basis of their specific characteristics. Statistical expectations for the frequency of repeated sequences were calculated on the basis of their nucleotide composition, their length and the length of segment I or II. On the other hand, 8 random sequences having the same nucleotide composition and the same length of sequece I were computer-generated and assessed for the frequency of repeats and palindromes. These sequences were used to check the correctness of statistical expectations and to evaluate the significance of the divergence of observed and expected frequencies of repeats in segment I, in view of the existence of statistical fluctuations.

3. Results and discussion

The analysis reported here is limited to: (a) direct repeats; (b) inverted repeats; (c) palindromes. As for repeats, only data on sequences longer than 12 nucleotides will be reported. This is justified by two reasons:

(1) So far, no excision site shorter than 13 nucleotides with one mismatch has been reported; this might correspond to the minimal length of sequences involved in site-specific illegitimate recombination;

(2) This sequence length is in the neighborhood of the longest direct repeats found in random sequences.

For palindromes, only data on sequences longer than 15 nucleotides will be presented.

Data on length, sequence and positions of repeats and palindromes of segments I and II are given in tables 1–3; (positions give the number of the first

```
          1                   25                  50                  75                 100
I.  CCGAAACTAAATAAAATATATTATTAATAATATTATATAATATAATAATAATATAATAATTTTATATAAATATATATTTATATATTAAATTAAATTATAA
    TTTTATTATGAAAATTATATCTTTTATTATATTCTCCTTTCGGGGTTCCGGCTCCGGCCGGAAAATCCCTCCCAGTCCTGACGGGGCTTGCTCACATCCT
    TATAATTAAATAAAGGTGCCAATATATTTTATATATAAATTAATATATATATATTATATATATATATACTTTTTTTTAATAATATATCTATATATATAAA
    TTAATATATTATATTATATTTTTATATAATATATTATTAATTATTATTTTAATTTTCTATTCTATTGTGGGGGTCCCAATTATTATTTTCAAAAATAATT
    ATTATTGGGACCCGGATATCTTCTTGTTTATCATTTATTATTTTATTAAATTTATTATTATTTTTAATTTATATTTATATTATATAATTAATTATATCGT
    TTATACCTTATTATTATTATATAATATATTATATTATTATAATATATTTATTGATTATATTATAAAATTTATTCTATGTGTGTTCTATATATATTTAATA
    TTCTGGTTATTGATCACCCACCCCCTCCCCCTATAAAACTTAGTTTATTACTTATATATTTATAAATATAAATCTAACTTAATTAATAATTTAAATAATA
    TACTTAATATTTAATAAATAGGAATAATTGTAACCTTTTTATAATTATATATAATAATAATATATTATCAAATAATTATTATTTCTTTTTTTTCTTTA
    ATTAATAATTAATTAATTAATTAATTAATATTTTATTTTATAAAAAATATATTTCTCCTTACGGGGTTCCGGCTCCGTAGCGGG

II. CCGGTTGTTCACCGGATTGGTCCCGCGGGGAATATTAATAATAAATTACAACATTTAAATAATATAAATAATTGAAATCTACAAATTTATAATTATAATA
    AAAATATAGAAATTATAAATACTATAAATGATAAATTAATTAATAAATTATTATATAAAATAATAACTTTAAAATTAAATAATATAAATATTAATAAAAT
    TATTATAAGTAAACTTATTAATTAACACAGTTTAAATAAATTAAATATTAAATTTATTATTATAATAATGATATTAATAATAATAATAATAATAATAATA
    ATAATTATTATATAAATATAATAAATAAATTAATAAATATTATAAATAATAATATAAATAATAATTTATGTAATATTTTAAGTTATTATTATAATAAAAA
    AGTAACTATTGAACCTATTAAATTATCATATATTTATTTAAATAGTGATATTTTTAGTAAATATATTAGTTTAAATGATATAGATAAATATAATAATGGT
    ATCTTAACTAATTATCAACGTATATTAAATAATATTATGCCTAAATTAAATGATCAACCAACTAAATATTTATATATTAAAATATCTATAGGAATATTAT
    CAATAGTTATATTATTATAAATATTATTAATATTACCAATATAATTATTATTATTATTATTATAATTATTATTATTAATATTATTATTATTATTATT
    ATTATTATTATTATTTAATAAATTAATTATATTATTATATTTATTATTATTAATATTATTAATATTATTAATATAATTTATAGAAATATTATGATCTATT
    AAATTTAAAGGTAGTTTAAGTAATAATAATGGTAGAACTAGTACACTTAATTTATTAAATGGTACTTTTAATAATAAAAAATATTTATGAAGTAATATTA
    ATAATAATTATAAATTAAATTATATCCCTTCTAATCATAATTTATATAATAATTCTAATATTAATAAAAATGGTAAATATAATATTAAAGTTAAATTAAA
    CTTTATTTAATATATATATTAATAGTCCGG
```

Fig.1. Primary structure of segments I and II (see text).

nucleotide in the repeats or palindromes of fig.1). A few remarks should be made here:

(i) Simple repeats are sequences which exist (in the same or in the reverse nucleotide order) at two or more non-overlapping sites (fig.2a); they are listed in tables 1,2; shorter sequences internal to these repeats (or to palindromes) and their inverted repeats are not listed; this explains why the vast majority of repeats listed have a frequency of only 2;

(ii) Overlapping repeats (fig.2b) are sequences usually formed by very short internal repeats; they are not listed;

(iii) Shared sequences (fig.2c) are often found in repeats and palindromes; they are indicated in tables 1–3 by dots indexing the position numbers; inverted repeats included in palindromes are not listed.

The sequences of tables 1–3 are exclusively made up of A and T, with only 3 remarkable exceptions

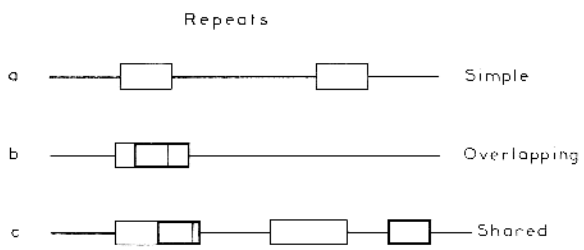

Fig.2. Scheme of simple, overlapping and shared repeats (see text).

Table 1
Direct repeats

	Length	Sequence	Positions	
I.	12	TAATTATTATTT	338*	775*
	"	AATTATTATTTT	339*	378
	"	TATATATTTATA	71	653
	"	ATATTATATAAT	30*	477
	"	ATATATATATAT	243	256
	13	TTAATTAATAATT	679	798
	14	AAATAATTATTATT	393	772*
	"	TATTATATAATATA	31*	515*
	"	AATATATTATTAAT	15	328*
	16	TAATATATTATATTAT	302	522*
	"	TTATATAATATATTAT	322*	517*
	"	ATATAATAATAATATA	40*	749
	"	CGGGGTTCCGGCTCCG	141	862
	19	TATATATAAATTAATATAT	230	291
II.	12	TAAATAAATTAA	233	322*
	"	ATAAATTAATTA	131	718*
	"	ATTAAATAATAT	174*	524
	"	TAATATTAATAA	893*	956*
	"	ATTTATTATTAT	252*	739*
	"	ATAATAATGGTA	490*	823
	13	ATAAATATAATAA	312*	484*
	"	AATATTAATAAAA	187	957*
	"	ATTATAATAAAAA	92	388*
	"	TAATAATAATTAT	296*	899*
	"	TTATATTATTATA	607	727*
	14	AATATTAATAATAA	31	894*
	"	AATTATTATATAAA	146	303*
	"	ATATTAATAATAAT	271	895*
	15	AATAATATAAATAAT	58*	348
	"	TTATTATTATAATAA	254*	383*
	16	TTAAATAATATAAATA	55*	175*
	"	AATATTATTAATATTA	620	752*
	20	TATAATTATTATTATTATTA	640	661*
	21	TTATTATTATTAATATTATTA	669*	741*

(2 direct repeats and 1 palindrome). They fully confirm our previous conclusion that yeast mitochondrial DNA contains 'reiterative sequences involving alternating and non-alternating AT' [9] as well as our prediction 'on statistical grounds alone, of the existence in the spacers of a large number of homologous sequences long enough to allow stable duplexes to be formed by a recombination mechanism' [10]. In fact the observed frequencies of, for instance, direct repeats found in segment I or II exceed those statistically expected for sequences ranging from 12–20 nucleotides (the GC-rich sequence 16 nucleotides long of table 1 having an exceptionally low statistical expectation). Most of the observed frequencies also are much higher than those occurring in random sequences. It should be pointed out that, neglecting shared sequences, direct and inverted repeats longer than 12 nucleotides each represent ~1/3 of all nucleotides in segments I or II; palindromes longer than 15 nucleotides represent ~1/5 of all nucleotides. (The rest of the segments are formed by shorter repeats and palindromes.) In contrast, direct repeats longer than 12 nucleotides only represent ~1/20 of all nucleotides in random sequences.

Table 2
Inverted repeats

	Length	Sequence	Positions			
I.	12	TTTATATAATAT	321	538		
	"	TATATTATATTT	310	474		
	"	TAATTTTATTAT	98	342		
	"	TATAATTAAATA	201	235*	296*	
	"	TATAATTAATTA	483*	820		
	"	ATTATTATATAA	513*	15*	328*	
	"	TATATTATATAA	476*	303*	523*	
	13	TATTATATTATAT	307	525		
	"	TTTTAATAATATA	274	51*		
	"	TAAATATATATTT	67*	228*		
	"	TTAATAATTAATT	802	679		
	"	TTATATATTTATA	652	74*		
	14	ATAATAATTATTAT	392	19*		
	15	ATATTATATAATATA	30*	519*		
	16	ATATAATAATAATATA	40*	749		
	17	AATTAATATATTATATT	299	475		
	18	ATTATTATATAATATATT	513	322		
II.	12	TTTAATAATAAA	867	356*		
	"	AAATAATATTAT	527	387		
	"	AATATAATAAAT	315*	56		
	"	AATATTATTATT	680	254*	383*	654*
	"	ATAATTATTATA	301*	621	753	762*
	"	TAAATAATAATT	355*	35*		
	13	AATAATTATAAAT	903	185		
	"	AATTTATTATTAT	251	706		
	"	ATATTATTATATT	729*	607		
	14	TAATAATATAAATA	347	484		
	"	AATATAATTATTAT	638	763		
	"	ATTTATTATTATTA	739*	704*		
	"	AATAATAATTATAA	900*	31		
	15	AATTATAAATTAAAT	906*	237*		
	16	TATAAATATAATAAAT	311*	176*		
	"	TATTAATAATAATAAT	272	293		
	"	TATTATAATTATTATT	658*	672*	744*	

Two additional points, now under investigation, have led to some preliminary conclusions:

Table 3
Palindromes

	Length	Sequence	Position
I.	15	TAATAATATAATAAT	46*
	"	TATATATTTATATAT	71*
	"	TATATTATATTATAT	305
	16	TATTATTAATTATTAT	332
	17	AATATTTTATTTTATAA	827*
	"	TTAATTAATAATTAATT	798*
	18	ATTTCTTTTTTTTCTTTA	783*
	"	ATATTAAATTAAATTATA	82*
	"	TATTATATAATATATTAT	515
	22	ATAATATAATAATAATATAATA	37*
	24	ATATATATATATTATATATATATA	243
	26	ATTATATATAATAATAATATATTATTA	744
	"	ATAATTAATTAATTAATTAATTAATA	805*
II.	15	TTATATTATTATATT	727
	19	ATTATTATAAATATTATTA	611
	20	TAATAAATATTATAAATAAT	331
	"	AATATTAATAATAATTATAA	894
	33	TATTATTATTATTATTATTATTATTATTATTAT	682
	35	ATTAATAATAATAATAATAATAATAATAATAATTA	273
	58	ATTATTATTATTATTATTATAATTATTAT TATTATTAATATTATTATTATTATTATTA	644*

1. If 5% mismatch is allowed (a value lower than that found at two excision sites [4]) the number (or the length) of the repeats is greatly increased over the values of tables 1–3; for instance, in segment I, neglecting overlapping repeats, there is no direct repeat longer than 19 nucleotides, but three pairs of repeats 20 nucleotides long are found if a single mismatch is allowed; furthermore, each of these repeats (137,858; 318,513; 228,289) is found in several overlapping frames.
2. A comparison of segments I and II (which might, however, be on different strands) has revealed that many sequences of segment I are repeated in segment II, in agreement with the idea that spacer sequences are built according to the same pattern all over mitochondrial genome units; this idea is also supported by the similar level of repeats found in the two segments (see tables 1–3), and by previous experiments showing similar pyrimidine isostich patterns and similar GC vs. ρ (buoyant density) relationships for 'petite' genomes arising from different regions of the wild-type genome [11].

No exception apparently exists to this rule, since even the 68 nucleotide segment forming the repeat units of mitochondrial DNA of petite RD1A [12], though claimed to be 'unique' [13], shares several sequences 14–18 nucleotides long with segments I and II; since these segments only represent <5% of all spacers, many more such (and longer) sequences must be present in the mitochondrial genome.

This work fully substantiates our suggestions, opposed [14,15], that the mitochondrial genome of yeast is highly repetitious in nucleotide sequences in its AT spacers and GC clusters. The abundance of such repetitive sequences appear to account, as predicted [2,16,17], for the extremely high frequency of the spontaneous 'petite' mutation and of mitochondrial recombination in crosses, as well as for the apparent total sequence homology [18] of mitochondrial genomes, which originate from different strains and differ in the lengths of their AT spacers.

A more detailed analysis of segments I and II and of other spacer–cluster sequences will be presented elsewhere.

References

[1] Barnardi, G., Carnevali, F., Nicolaieff, A., Piperno, G. and Tecce, G. (1968) J. Mol. Biol. 37, 493–505.
[2] Bernardi, G. (1979) Trends Biochem. Sci. 4, 197–201.
[3] Faugeron-Fonty, G., Culard, F., Baldacci, G., Goursot, R., Prunell, A. and Bernardi, G. (1979) J. Mol. Biol. 134, 493–537.
[4] Gaillard, C., Strauss, F. and Bernardi, G. (1980) Nature 283, 218–220.
[5] Bernardi, G., Baldacci, G., Culard, F., Faugeron-Fonty, G., Gaillard, C., Goursot, R., Strauss, F. and De Zamaroczy, M. (1980) submitted.
[6] Baldacci, G., De Zamaroczy, M. and Bernardi, G. (1980) FEBS Lett. 114, 234–236.
[7] De Zamaroczy, M., Baldacci, G. and Bernardi, G. (1979) FEBS Lett. 108, 429–432.
[8] Tzagoloff, A., Macino, G., Nobrega, M. P. and Li, M. (1979) in: Extrachromosomal DNA (Cummings, D. J. et al. eds) pp. 339–355, Academic Press, New York.
[9] Bernardi, G., Faurès, M., Piperno, G. and Slonimski, P. P. (1970) J. Mol. Biol. 48, 23–42.
[10] Prunell, A. and Bernardi, G. (1974) J. Mol. Biol. 86, 825–841.
[11] Faugeron-Fonty, G. (1978) PhD Thesis, Université Paris VII.
[12] Van Kreijl, C. F. and Bos, J. L. (1977) Nucleic Acids Res. 4, 2369–2388.
[13] Bos, J. L., Van Kreijl, C. F., Ploegaert, F. H., Mol. J. N. M. and Borst, P. (1978) Nucleic Acids Res. 5, 4563–4578.
[14] Borst, P. (1972) Ann. Rev. Biochem. 41, 333–376.
[15] Borst, P. and Grivell, L. A. (1978) Cell 15, 702–753.
[16] Fonty, G., Goursot, R., Wilkie, D. and Bernardi, G. (1978) J. Mol. Biol. 119, 213–235.
[17] Prunell, A., Kopecka, H., Strauss, F. and Bernardi, G. (1977) J. Mol. Biol. 110, 17–52.
[18] Groot, G. S. P., Flavell, R. A. and Sanders, J. P. M. (1975) Biochim. Biophys. Acta 378, 186–194.