

## A new approach to the study of nucleotide sequences in DNAs

---

Stanislav D. Ehrlich, Jean-Paul Thiery, Anne Devillers-Thiery  
and Giorgio Bernardi

---

Laboratoire de Génétique Moléculaire, Institut de Biologie Moléculaire,  
Paris 5<sup>e</sup>, France

---

Received 6 November 1973

---

### ABSTRACT

The composition of the 3' terminal, 5' terminal and 5' penultimate nucleotides of the oligonucleotides released by spleen acid DNase and snail acid DNase from five 'repetitive' DNAs (guinea pig, mouse and crab satellite DNAs, yeast mitochondrial DNA and poly (dAT : dAT)) and four 'eukaryotic' DNAs (calf thymus, guinea pig and mouse liver DNAs, and yeast nuclear DNA) have been investigated and found to deviate in characteristic ways from those expected for bacterial DNAs having comparable base compositions. The deviation patterns obtained represent a novel way of characterizing and comparing different DNAs on the basis of the frequency of the nucleotide sequences they contain.

### INTRODUCTION

We have recently shown that deoxyribonucleases (DNases) hydrolyze specific sets of short nucleotide sequences and therefore can be used to obtain information on the frequency of such sequences<sup>1</sup>. Using methods described elsewhere<sup>2</sup> for the isolation and analysis of the termini, i.e., the nucleotides near the breaks introduced by the enzymes,  $WX^{\downarrow}YZ$ , (the sequence being written in the usual 5'-3' direction, and the vertical arrow indicating the position of the break), it is possible to show that the base composition of termini: (a) differs from the values expected for random degradation, in which case the composition of each terminus should be equal to the average base composition of the DNA; (b) differs according to the enzyme used, indicating that different sets of sequences are split in a given DNA by different enzymes; (c) does not vary, as a rule, according to the level of DNA degradation. The minimum length of the sequences recognized by the DNases we have investigated so far is four nucleotides for the hog spleen acid DNase<sup>3-6</sup>, two

nucleotides for the snail hepatopancreas acid DNase<sup>7,8</sup> and three nucleotides for bovine pancreatic DNase<sup>9</sup> and for E.coli endonuclease I<sup>10</sup> (only positions XYZ were analyzed for the latter three enzymes).

Since DNases split specific sets of sequences, the base composition of termini is basically related to the relative amount of such sequences in the DNA under consideration. This is shown by the fact that the composition of termini released by the same DNase from DNAs having different G+C contents are different. If the compositions of termini released from bacterial DNAs are plotted against their G+C contents, linear relationships are obtained. The choice of bacterial DNAs in order to establish such relationships is justified (a) by the fact that bacterial DNAs do not contain short repetitive sequences, and (b) by the fact that the doublet frequencies of bacterial DNAs, as determined by the nearest neighbor analysis, show essentially linear relationships with the frequencies predicted for random association<sup>11</sup>, indicating a common type of doublet distribution in these DNAs.

If the DNAs under investigation contain short repetitive sequences, it can be expected that the base composition of termini released from them in general will deviate, in either direction, from that expected for non-repetitive (bacterial) DNAs having the same G+C contents. We show here that such is the case for 'repetitive' DNAs (as we will indicate the guinea pig, mouse and crab satellite DNAs, the yeast mitochondrial DNA and poly (dAT:dAT)) and for 'eukaryotic' DNAs (calf thymus, guinea pig and mouse liver DNAs, and yeast nuclear DNA).

A preliminary report on part of these investigations was presented elsewhere, together with a fuller discussion of the problem of DNase specificity<sup>1</sup>.

#### MATERIALS AND METHODS

Mitochondrial DNA from wild-type Saccharomyces cerevisiae cells, Cancer pagurus (A+T)-rich satellite DNA, guinea pig  $\alpha$ -satellite DNA and mouse satellite DNA were preparations obtained according to methods already described 12-14.

Poly(dAT:dAT) was obtained from Miles, Elkhart, Ind.

Calf thymus, mouse liver and guinea pig liver DNA were prepared according to the detergent procedure<sup>15</sup> and shown by hydroxyapatite chromatography to be free of material eluting before DNA (mono- and oligonucleotides and RNA).

Yeast nuclear DNA was prepared from the S.cerevisiae cytoplasmic 'petite' strain DM<sub>I</sub><sup>16</sup>, as described elsewhere<sup>12</sup>.

Other materials and methods were already described<sup>2,6-10</sup>.

### RESULTS AND DISCUSSION

The base compositions of the 3' terminal, 5' terminal and 5' penultimate nucleotides released from 'repetitive' DNAs by hog spleen acid DNase and Helix aspersa hepatopancreas acid DNase are presented in Tables I and II, respectively.

Table I : Termini released by spleen acid DNase from "repetitive" DNAs

DNA (G+C %)	3' terminal				5' terminal				5'penultimate			
	T	G	A	C	T	G	A	C	T	G	A	C
Poly (dAT) (0 %)	22	0	78	0	78	0	22	0	22	0	78	0
Crab (3 %)	26	3	69	2	72	2	20	5	17	3	80	0
Yeast (18 %)	31	18	46	5	33	26	29	13	21	15	59	5
Mouse (36 %)	22	28	37	13	26	28	26	20	21	17	58	4
Guinea pig (40 %)	27	38	28	7	10	56	18	16	11	46	35	8

Table II : Termini released by Helix acid DNase from "repetitive" DNAs

DNA (G+C %)	3' terminal				5' terminal				5'penultimate			
	T	G	A	C	T	G	A	C	T	G	A	C
Poly (dAT) (0 %)	13	0	87	0	87	0	13	0	13	0	87	0
Yeast (18 %)	15	5	78	1	18	40	28	14	37	12	44	17
Mouse (36 %)	18	3	77	2			n.d.				n.d.	
Guinea pig (40 %)	16	10	73	2	12	44	18	26	26	30	20	23

In all cases, except for poly (dAT:dAT), analyses of the 3' terminal nucleotides as released at different degradation levels by spleen acid DNase showed, above average sizes of the digest equal to 10 nucleotides, trends already described<sup>6</sup>, involving an increase of G and a decrease of the other three nucleotides.

The data reported in the Tables refer to average sizes of the digests equal to 15 and 18 for spleen acid DNase and for snail acid DNase, respectively.

Fig. 1 shows, as an example, a plot of the 5' terminal nucleotides, as released by the spleen enzyme, versus their G+C contents. The experimental values generally show strong deviations, in either directions, from the linear relationships found for bacterial DNAs, which are also presented in the figure.

Figs. 2 and 3 show the results of Tables I and II in the convenient form of deviation patterns. The histograms show the differences found between the base compositions of the termini released from 'repetitive' DNAs and those expected for bacterial DNAs having the same G + C contents.

The results obtained in the analysis of termini of 'repetitive' DNAs indicate that a number of short nucleotide sequences, recognized and split by the enzymes in these DNAs, are present in amounts which are larger or smaller than those present in bacterial DNAs of comparable G + C contents. Such different amounts of short sequences are the source of the different compositions of termini shown in fig. 1 and originate the deviation plots of fig. 2 and 3. The results obtained were expected since in all cases the DNAs examined were known to contain a large number of short repeated sequences. It may be interesting to remark that deviation patterns obtained with the same DNA as degraded by different enzymes are different, since different sets of short sequences are split, and that deviation patterns obtained with different DNAs as degraded by the same enzyme are also different since the frequency of the short sequences split by the enzyme is different in different DNAs.

Table III gives the base composition of the 3' terminal, 5' terminal and 5' penultimate nucleotides released from eukaryotic DNAs by the spleen enzyme and the composition of the 3' terminal nucleotides as released from the same DNAs by the snail DNase.

Fig. 4 shows the data of Table III in the form of deviation patterns; as in the case of figs. 2 and 3, the histogram presents the differences between the compositions of termini as released from eukaryotic DNAs and those expected for bacterial DNAs having the same G + C

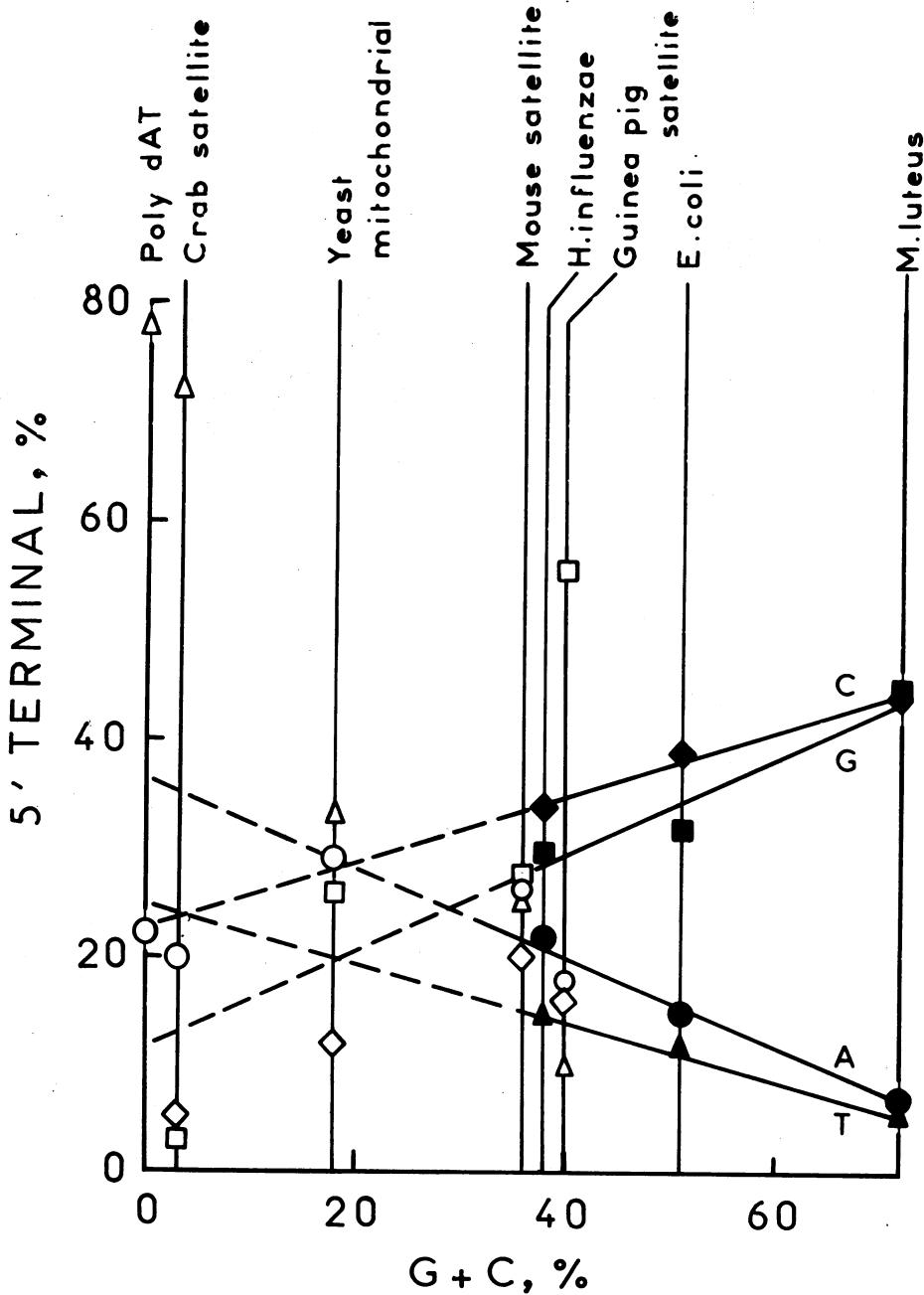


Fig. 1.

Composition of 5' terminal nucleotides as released by spleen acid DNase from 'repetitive' DNAs as plotted against their G + C contents (open symbols). Data for bacterial DNAs are also shown (filled symbols).

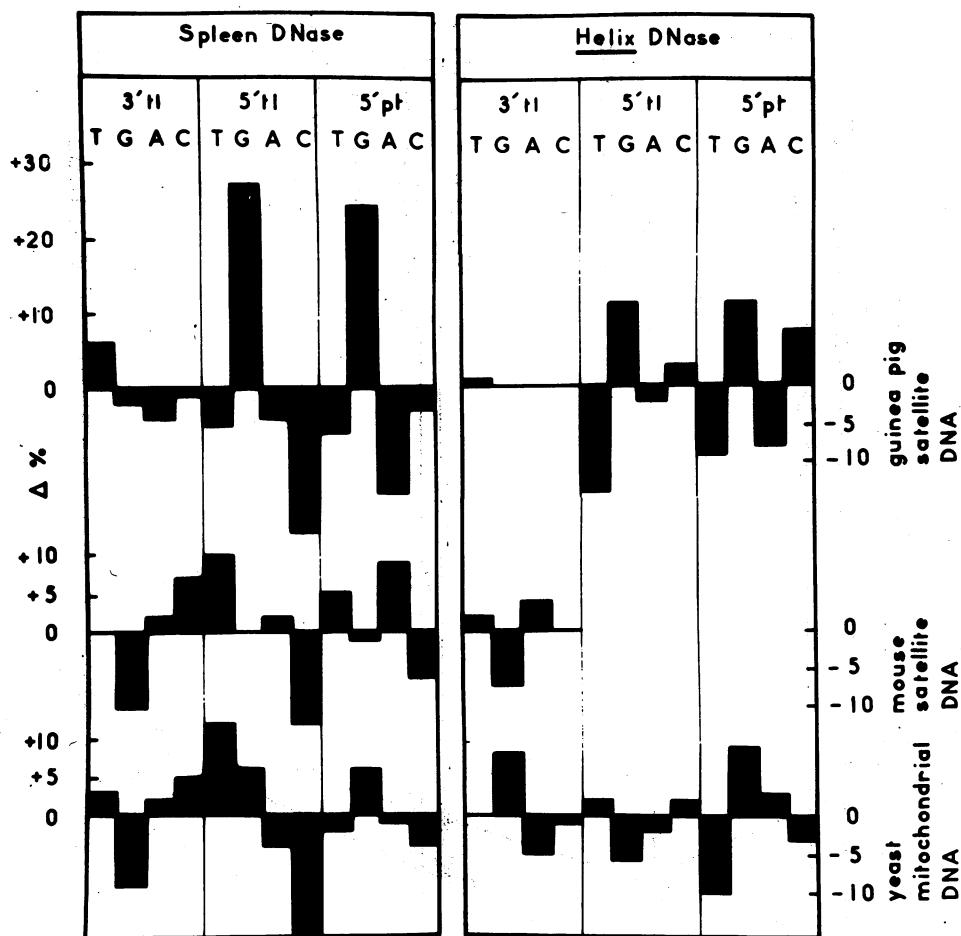


Fig. 2.

Deviation patterns of three 'repetitive' DNAs. The histograms show the differences between the composition of termini formed from guinea pig satellite, mouse satellite and yeast mitochondrial DNAs by spleen and snail DNase and the compositions expected for bacterial DNAs having the same G + C contents, values represent differences in the percentages of each terminus.

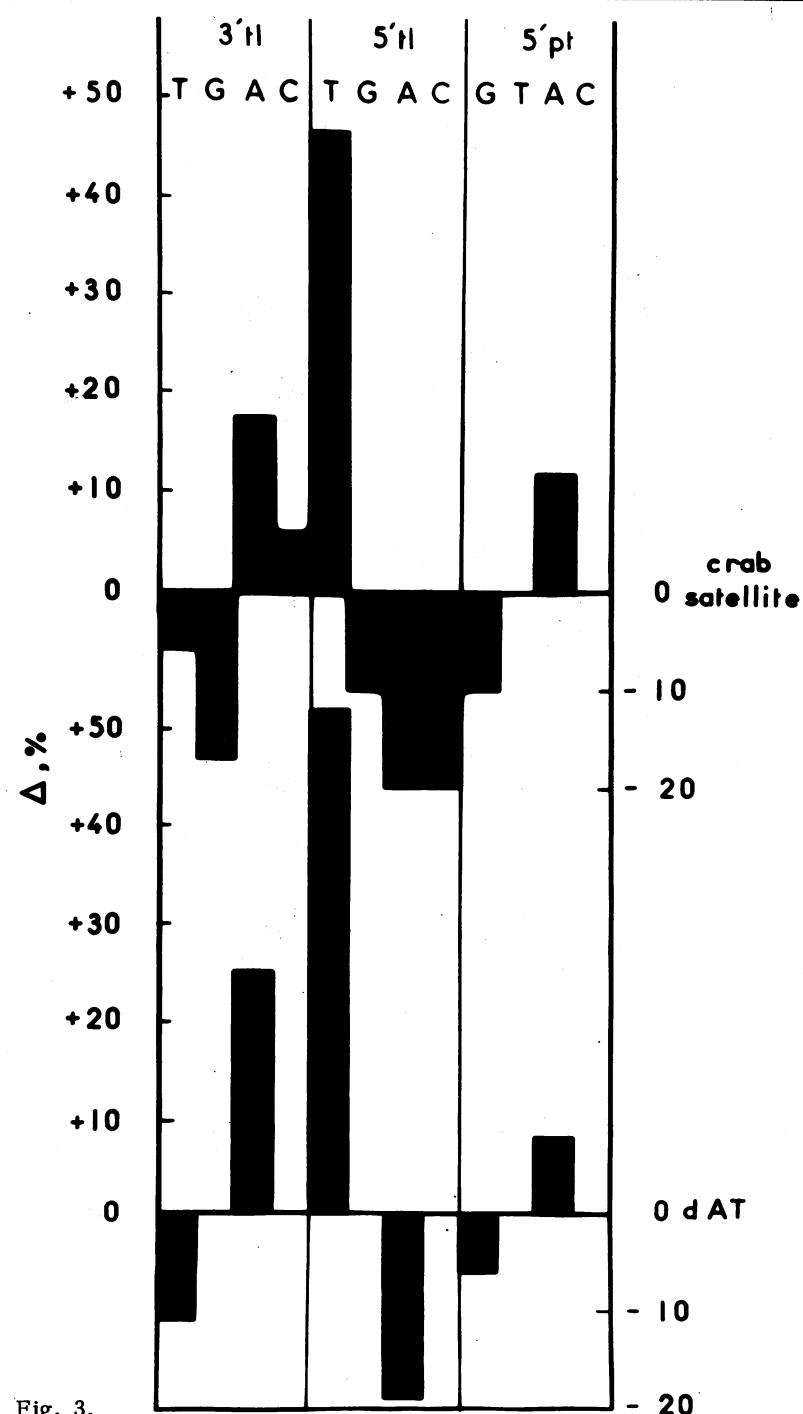


Fig. 3.

Deviation patterns of crab satellite DNA and poly (dAT:dAT) as obtained after spleen acid DNase digestion. See legend of fig. 2 for other indications.

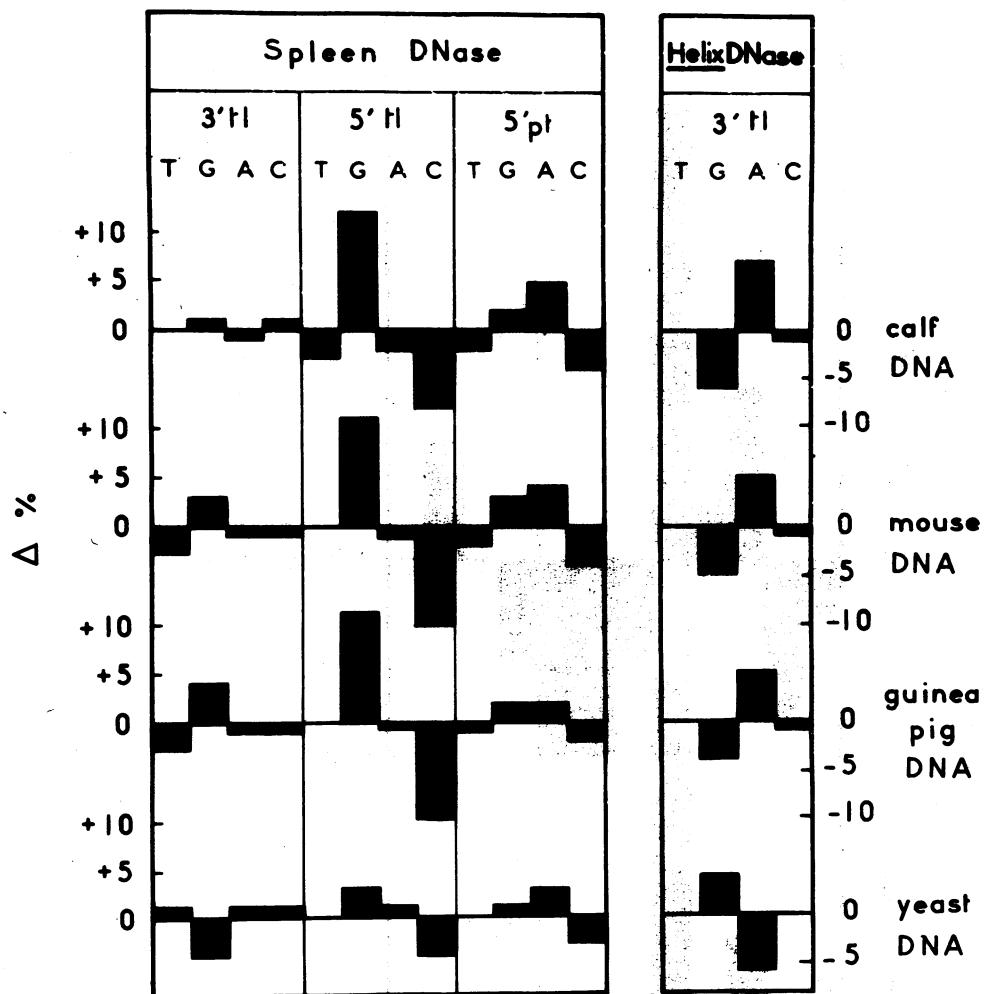


Fig. 4.

Deviation patterns of four 'eukaryotic' DNAs. See legend of fig. 2 for other indications.

Table III :Termini released by spleen and snail DNase from four "eukaryotic" DNAs.

DNA (G+C %)	3' terminal				Spleen DNase				5' penultimate				Snail DNase			
	T	G	A	C	T	G	A	C	T	G	A	C	T	G	A	C
Yeast (39 %)	23	35	33	8	15	31	24	30	17	22	53	8	16	13	68	2
Guinea pig (41 %)	18	45	30	7	12	40	19	29	15	24	51	10	16	6	77	1
Mouse (42 %)	18	44	30	7	14	41	20	25	14	26	52	8	16	6	77	1
Calf (44 %)	20	43	29	8	11	43	18	28	14	26	52	8	16	6	78	1

contents. The deviation patterns of the three mammalian DNAs are very similar to each other; the most pronounced deviation concern the 5' terminal G and C in the spleen DNase digests. The deviation pattern of yeast nuclear DNA is different from those of mammalian DNAs and is characterized by smaller deviations from the relationships of bacterial DNAs.

The results obtained with 'eukaryotic' DNAs lead to a number of conclusions: (a) a number of short sequences are present in larger or smaller amounts in the 'eukaryotic' DNAs examined compared to bacterial DNAs of comparable G + C contents; (b) mammalian DNAs have very similar deviation patterns, whereas yeast nuclear DNA has a different deviation pattern characterized by lesser deviations from bacterial DNAs; (c) as for the origin of the deviations observed in mammalian DNAs, they evidently do not arise from the repetitive sequences of the satellite DNAs they contain, since the deviation patterns of mouse and guinea pig satellite are very different from each other and from the corresponding total mammalian DNAs (compare fig. 2 and fig. 4); the deviations must therefore originate in nucleotide sequences contained in the main ( $1 \cdot 697 \text{ g/cm}^3$ ) DNA component and/or in components similar to the  $1 \cdot 704$  and  $1 \cdot 709 \text{ g/cm}^3$  components observed in the bovine genome<sup>17</sup>; such sequences must share common features in mammalian DNAs and also must be abundant, as suggested by the fact that the repetitive sequences of satellite DNAs, which only represent 5-10% of genome in the case of mouse and guinea pig, cannot be recognized in the deviation patterns of total mammalian DNAs.

In conclusion, the deviation patterns obtained with DNases represent a novel way of characterizing and comparing different DNAs on the basis of the frequency of the nucleotide sequences they contain. The results obtained so far parallel those obtained by another frequency method, the nearest neighbor analysis<sup>18,19</sup>; in fact, they extend the nearest neighbor data in that they concern sequences longer than dinucleotides<sup>1</sup>.

## FOOTNOTE

/ Present address: Department of Genetics, Stanford University Medical School, Stanford, California 94305.

## ACKNOWLEDGEMENTS

This work was supported, in part, by the Delegation Generale a la Recherche Scientifique et Technique (Comite Interactions Moleculaires en Biologie).

## REFERENCES

1. Bernardi G., Ehrlich S.D., Thiery J.P. (1973) Nature, New Biol., 246 36-40
2. Bernardi G., Ehrlich S.D., Thiery J.P. (1973) in Methods in Enzymology, (L. Grossman and K. Moldave, eds.), Academic Press, New York, in press
3. Carrara M., Bernardi G. (1968) Biochemistry, 7, 1121-1132
4. Ehrlich S.D., Torti G., Bernardi G. (1971) Biochemistry, 10, 2000-2009
5. Devillers-Thiery A., Ehrlich S.D., Bernardi G. (1973) Eur.J.Biochem., 38, 416-422
6. Thiery J.P., Ehrlich S.D., Devillers-Thiery A., Bernardi G. (1973) Eur.J. Biochem., 38, 434-442
7. Laval J., Thiery J.P., Ehrlich S.D., Paoletti C., Bernardi G. (1973) Eur.J. Biochem., in press
8. Ehrlich S.D., Devillers-Thiery A., Bernardi G. (1973) Eur.J.Biochem., in press
9. Ehrlich S.D., Bertazzoni U., Bernardi G. (1973) Eur.J.Biochem., in press
10. Ehrlich S.D., Bertazzoni U., Bernardi G. (1973) Eur.J.Biochem., in press
11. Kaiser A.D., Baldwin R.L. (1962) J.Mol.Biol., 4, 418-419
12. Bernardi G., Piperno G., Fonty G. (1972) J.Mol.Biol., 65, 173-189
13. Ehrlich S.D., Thiery J.P., Bernardi G. (1973) Biochim.Biophys.Acta, 312, 633-636
14. Corneo G., Ginelli G., Soave C., Bernardi G. (1968) Biochemistry, 7, 4373-4379
15. Kay, E.R.M., Simmons N.S., Dounce A.L. (1952) J.Am.Chem.Soc., 74, 1742-1726
16. Bernardi G., Carnevali F., Nicolaieff A., Piperno G., Tecce G. (1968) J.Mol. Biol., 37, 493-505
17. Filipski J., Thiery J.P., Bernardi G. (1973) J.Mol.Biol., 80, 177-197
18. Swartz M.N., Trautner T.A., Kornberg A. (1962) J.Biol.Chem., 237, 1961-1967
19. Subak-Sharpe J.H. (1969) in Handbook of Molecular Cytology, p.67-87 (A. Lima-de-Faria, ed.), North-Holland, Amsterdam