Palindromic units are part of a new bacterial interspersed mosaic element (BIME)

Eric Gilson, William Saurin, David Perrin, Sophie Bachellier and Maurice Hofnung*
Unité de Programmation Moléculaire et Toxicologie Génétique, CNRS UA271, Institut Pasteur, 28 rue du Docteur Roux, Paris 75015, France

Received February 12, 1991; Revised and Accepted March 13, 1991

ABSTRACT

Palindromic Units (PU or REP) were defined as DNA sequences of 40 nucleotides highly repeated on the genome of *Escherichia coli* and other *Enterobacteriaceae*. PU are found in clusters of up to six occurrences always localized in extragenic regions. By sorting the DNA sequences of the known PU containing regions into different classes, we show here for the first time that, besides the PU themselves, each PU clusters contains a number of other conserved sequence motifs. Seven such motifs were identified with the present list of PU regions. Remarkably, each PU cluster is exclusively composed of a mosaic combination of PU and of these other sequence motifs. We demonstrate directly by hybridization experiments that one of these motifs (called L) is indeed present at a large number of copies on the *Escherichia coli* chromosome and that its distribution follows the same species specificity as PU sequences themselves. We propose that the mosaic pattern of motif combination in PU clusters reveals a new type of bacterial genetic element which we propose to call BIME for Bacterial Interspersed Mosaic Element. The *Escherichia coli* genome contains about 500 BIME.

INTRODUCTION

A Palindromic Unit (PU) was originally defined in 1982 as a palindromic DNA element present several times in four intergenic regions of bacterial operons (1). Two years later, an extensive search revealed that a large number of intergenic sequences in *Escherichia coli* and in *Salmonella enteritica* serotype Typhimurium were homologous to PU DNA, so that PU amounted to almost 1% of the genomic DNA (2,3). A number of laboratories kept the name PU family for these sequences (2,4), while others called them REP for Repetitive Extragenic Palindromes (5,6). It was latter shown that the consensus sequence from *Escherichia coli* PU (3 and upper part in fig. 1), differs slightly from that of *Salmonella enteritica* serotype Typhimurium PU (abbreviated *Salmonella* PU) (3); in particular, *Salmonella* PU contain an additional G between the positions 11 and 12 of the consensus shown in figure 1. The existence of these species-specific sequence variations was confirmed by hybridization experiments (7).

Because of its palindromic nature, part of the PU sequence (32 out of the 40 nucleotides of the consensus) could adopt a 'stem and loop' structure (nucleotides 4−35 in fig. 1), and evidences for the formation of this secondary structure at the level of RNA was obtained from *in vitro* studies (6). The fact that the palindrome is not perfect has an important consequence: the PU can be oriented and we distinguish the right end and the left end of PU (see details in fig.1). PU are found in clusters, from one to six occurrences, always outside structural genes. Within a cluster, it is remarkable that successive PU rigorously alternate in orientation (fig. 1).

Except for the PU themselves, no other sequences common to PU clusters have been described. Indeed, various alignements of all known PU clusters did not reveal other conserved sequences. In the present paper, we perform a detailed analysis of the DNA sequences corresponding to the present list of PU clusters and show that they can be sorted into classes which disclose a number of new conserved sequence motifs in addition to PU themselves. Moreover, each PU cluster is exclusively composed of PU and of these conserved motifs. We demonstrate directly by hybridization experiments that one of these motifs (called L) is indeed present at a large number of copies on the *Escherichia coli* chromosome and that its distribution follows the same species specificity as PU sequences themselves. We propose that the mosaic pattern of motifs combination in PU clusters reveals a new type of bacterial genetic element, a Bacterial Interspersed Mosaic Element (abbreviated as BIME).

MATERIALS AND METHODS

Collection of *Escherichia coli* PU clusters

We defined a PU cluster as a DNA sequence that contains at least one occurrence of a typical PU sequence. In order to collect the sequence of most of the known *Escherichia coli* PU clusters, we used two different computer programs: PROBE3 (15) and FASTN (16). Both programs were used to search Genbank with different probes (release 58.0; 24.600.000 bases in 21.248 sequences).

* To whom correspondence should be addressed

The PROBE3 program was used to search the databank for sequences similar to two complementary probes S1 and S'1. Both sub-sequences of S1 and S1' correspond to the most conserved parts of the PU consensus sequence (fig. 1). Sequence S1 is described as 5'-GCC2GATGCGR(0,20)TTATC8GGCCTA-CR-3' where R stand for purine, Y for pyrimidine, 2 for T or G, 8 for A or C and (0,20) means that both sub-sequences of each probe are separated by at most 20 bases (15). Sequences were kept only if they had at least 81% residues matching the 5'-subsequence of S1 and 78% residues matching the 3'-subsequence of S1 (Search A in Table I). The requirement in the search for S'1 was 78% matches with 3'-subsequence and 81% matches with 5'-subsequence (Search B in Table I). A third set of sequences was obtained using program FASTN. In this case, we searched the sequence databank for sequences homologous to the string: 5'-GCCGGATGCGGCGTAAACG-CCTTATCCGGCCTAC-3' (called S2). This string is homologous to the PU consensus sequence (fig. 1). The matches with a similarity score higher than 66 (16) were collected (search C in table I).

About 30% of the sequences we obtained from these three computer searches did not originate from *Escherichia coli* and were subsequently discarded. Finally we added some *Escherichia coli* PU sequences screened by eye (search E in table I).

The compilation of the computer and eye searches defined a set of 94 PU clusters with 172 PU occurrences, called the initial PU cluster set. The details of these sequences are given in table I. We added in Table I 19 recently described PU clusters which have not been subjected to the multiple sequence alignment analysis (see below).

'Multiple sequence alignment' procedure

In order to align sets of sequences we used an iterative procedure. At each step three successive tasks were performed: (i) all the sequences were independently aligned with a profile (17), providing pairwise alignments of the sequences with the profile, (ii) a multiple alignement of the sequences was built from the pairwise alignments of the profile with each sequence, (iii) a new profile was derived from this multiple alignment. This procedure was stopped when the new profile was the same as the one of the previous step. The initial profile was either derived from an alignment made by hand or was one sequence chosen at random in the set to be aligned. Pairwise alignments of a sequence with a profile was performed using a standard dynamic programming procedure (18). CONSALI, a C program running under UNIX, performs that task and is available upon request.

Sequence Variation Pattern and consensus sequence

In order to characterize the multiple alignments, we defined the Sequence Variation Pattern: it is basically a histogram of bases and gap frequencies at each position. We defined the gap frequency at position i (GF_i) as $100.\Delta_i/T$, where Δ_i is the number of gaps in the i^{th} column of the alignment and T the total number of sequences. We defined the effective frequency of base B at position i (denoted EF_{iB}, B being either A, T, G or C) as $100.N_{iB}/(T-\Delta_i)$ where N_{iB} is the number of base B in the i^{th} column. Positions where the gap frequency was higher than 60% were not taken into account in the Sequence Variation Pattern. The Sequence Variation Pattern is then defined as both the histogram of the bases of which effective frequency is higher than 25% and the histogram of gap frequencies. We defined the consensus sequences as the bases present in the Sequence

Table I. Compilation of known BIME.

For each PU cluster, we have presented, from left to right: its name; the type of search by which it was detected (searches A, B, C or E; see Materials and Methods); the number of the left external segments which is referred in fig. 2 (LES); the coordinates of the palindromic part of the PU, <...>> means a PU in the orientation shown in fig. 1. Each BIME can be described by a specific combination of segments (column BIME structure). For clarity, the PU sequences are indicated by a '<' or a '>' according to their orientation ('>' means a PU in the orientation shown in fig. 1). For example, the combination 'B>L<s>' defines the corresponding BIME as being composed of three PU and of three motifs located outside a PU (one B, one L, and one s); the order of the segments, from left to right, corresponds to their order at the DNA level, from 5' to 3'. The segments presented between [ ] means that they exhibit only a slight homology with the corresponding consensus; between () is indicated the name of the BIME if several different BIME are present in the same PU cluster referred in the first column as well as the name of some internal segments which are discussed in the text. The PU* have been described in (8).

NAMES	SEARCHES A B C E	LES	PU COORDINATES	BIME STRUCTURES
ECOALKA	+ + + -		< 1429 1458>>	>S<
		17	<< 1481 1512 >	
ECOAPAH	+ + - -		<< 2274 2306 >	<
ECOAPPA	- + - -		<< 294 325 >	<s
ECOARAABD	+ + + -		< 3356 3386>>	[S<s]>S<s>S<s>S<[s]e
			<< 3408 3432 >	
			< 3440 3471>>	
			<< 3493 3517 >	
			< 3525 3556>>	
			<< 3578 3602 >	
ECOARLACA	- + - -	58	<< 12 50 >	<
ECOAROG	- + - -	60	< 1569 1600>>	><6><B (=ECOAROG1)
		23	<< 1613 1643 >	
		59	<< 2067 2102 >	< (=ECOAROG2)
ECOASPC	+ - - -	1	< 1389 1413>>	B>L
ECOAVT[d]	- - - +		< 100 127 >>	A>L<B
			<< 171 195 >	
ECOCCA	+ + + -	54	< 1703 1732>>	>S<s>
			<< 1755 1786 >	
			< 1794 1823>>	
ECOCDHA	+ + + -	2	< 1112 1136>>	B>L<A
		41	<< 1180 1209 >	
ECOCHE3	- - + -	24	< 3010 3041>>	>
ECOCPDB	- + - -		<< 2107 2133 >	<s>S[<]
			< 2140 2166>>	
ECOCPX	+ + + -		<< 781 810 >	<1>
			< 824 856>>	
ECOCYD	+ - + -	42	< 191 220>>	A>L<B
		3	<< 264 290 >	
ECODEOCA	+ + - -	18	< 39 63>>	>S<
		55	<< 87 113 >	
ECOEXBBD[d]	- - - +		< 1758 1782>>	[<1]>L<1>L<[1>]e
			<< 1827 1856>	
			< 1871 1893>>	
			<< 1937 1966>	
ECOFDAPGK	- - - +	16	< 5545 5569>>	B>L<A
		53	<< 5613 5642>	
ECOFEPAA	+ - + -	43	< 2548 2577>>	A>L<B
			<< 2621 2645 >	
ECOFHUB	+ + + -	56	< 2213 2244>>	>S<
			< 2268 2297 >	
ECOFPG	- - - +		< 945 966 >	PU*
ECOGDHA	+ - - -	19	< 1546 1571>>	>L<r>
			<< 1614 1638 >	(r=r23)
			< 1668 1701>>	
ECOGLN	+ - - -		<< 1467 1483 >	<r>S[<]
			< 1509 1540>>	(r=r19)
ECOGLNHPQ[d]	- - - +		< 2948 2970>>	B>L<A
			<< 3015 3043 >	
ECOGLNSA	+ + + -		<< 1990 2019 >	[>L]<r1>L<r2[>L<]e
			< 2053 2077>>	(r1=r27a; r2=r26b)
			<< 2121 2150 >	
ECOGLTA	+ + + -		<< 6317 6348 >	s<s><10><(=ECOGLTA1)
			< 6356 6384>>	
		20	<< 6405 6429 >	
			<<10669 10698 >	<s>S<s>S<s[e]
			< 10706 10737>>	(=ECOGLTA2)
			<<10758 10786 >	
			< 10794 10826>>	
			<<10846 10871 >	
ECOGLYA	+ + + -	44	< 1651 1680>>	A>L<B[f]
			<< 1724 1748 >	
			< 1780 1806>>	
ECOGLYK	- - + -	61	< 1553 1581>>	>
ECOGUT	+ - - -	21	< 3263 3287>>	>
ECOHAG	+ - - -	4	< 1610 1634>>	B>L
ECOHISPUR	+ + + -	22	< 3866 3897>>	>S<s[>]
			<< 3920 3949 >	
ECOICDHKP	+ + + -		<< 317 345 >	S<4>S<1>
			< 360 391>>	
			<< 414 443 >	
			< 454 483>>	
ECOLAC	+ - - -	57	< 5673 5700>>	> (=ECOLAC1)
			< 6351 6375>>	>[L<] (=ECOLAC2)
ECOLACT	+ + + -	45	< 1242 1271>>	A>S<
		5	<< 1296 1320 >	
ECOLSP	+ + + -	25	< 939 970>>	>S<
		62	<< 995 1024 >	
ECOMALB	+ + + -		<< 1593 1622 >	<s>S<A
			< 1630 1661>>	
		63	<< 1685 1714 >	

ECOMALM	+ + + -		<< 14 43 >	<s>S<s[>]
			< 51 82>>	
			<< 105 134 >	
ECOMALT	- + - -	64	< 2894 2917>>	>
ECOMELB	- - - +	6	< 1495 1519>>	B>L<
			<< 1560 1575 >	
ECOMETJA	+ + + -	7	< 633 657>>	B>L<
			<< 700 729 >	
ECOMETL	- + - -	26	<< 1328 1352 >	<
ECOMETLB2	- + + -		<< 84 113 >	[>]L<r> (r=r31)
			< 151 173>>	
ECOMOTAB[a]	- + - -	27	<< 1813 1844 >	<B
ECOMTLA	+ - - -	28	< 2093 2117>>	>L<
		65	<< 2162 2189 >	
ECOMUTT	+ + + -		<< 554 583 >	[>L]<l>L<l>[e]
			< 597 628>>	
			<< 670 699 >	
			< 713 745>>	
ECONDH	+ + + -		<< 12 41 >	<l> (=ECONDH1)
			< 56 78>>	
			<< 1671 1700 >	<l> (=ECONDH2)
			< 1712 1740>>	
ECONRDA	+ + + -		<< 5836 5867 >	[>S]<s>S<>[S<][e]
			< 5875 5903>>	
			<< 5922 5953 >	
			< 5954 5984>>	
ECONTRLA	- - - +	49	< 1089 1117>>	A>L<B
		12	<< 1161 1185 >	
ECONUSA	+ + + -		<< 5052 5081 >	<s>
			< 5089 5120>>	
ECOP1PARA	- - + -	66	<< 3530 3560 >	<A
ECOP15B35[d]	- - - +		<< 102 131 >	<s>
			< 139 170>>	
ECOPGI[d]	- - - +		<< 2171 2201>	<
ECOPHOE	- - - +		< 1539 1559 >	PU*
ECOPHOS	+ + + -	68	<< 1332 1362 >	<A (=ECOPHOS1)
		67	< 3355 3384>>	>S<
		29	<< 3394 3418 >	(=ECOPHOS2)
ECOPHRORF	+ - + -	69	< 1991 2020>>	>
ECOPLSB	+ + + -	70	< 2990 3019>>	>L<lPU*
			<< 3064 3088 >	
			< 3102 3119 >	
ECOPOLA	+ + - -	30	< 3368 3392>>	>L<
		71	<< 3426 3446 >	
ECOPONA	- - + -		<< 2743 2764 >	<
ECOPROU[d]	- - - +		<< 4009 4039>	<l>
			< 4053 4084>>	
ECOPUREK	- - - +	86	< 1983 2011>>	A>S<l><
			<< 2037 2060 >	
			< 2075 2100>>	
		36	<< 2107 2139 >	
ECOPURF[d]	- - - +		< 104 135>>	>S<
			<< 158 187>	
ECORCSBC[d]	- - - +		<< 3481 3512>	<l>L<l>
			< 3526 3556>>	
			<< 3596 3625>	
			< 3639 3670>>	
ECORECD	- + - -	72	< 2104 2144>>	>
ECORECFA	+ + + -	8	< 3616 3640>>	B>L<
		46	<< 3684 3713 >	
ECOREPHEL	- + - -	73	<< 2443 2474 >	<
ECORF2X[a]	- - + -	74	<< 103 135 >	<
ECORPSRPO	- + - -	75	<< 4833 4861 >	<
ECOSODB	+ + - -	9	< 782 806>>	B>L<A
		47	<< 828 855 >	
ECOTGOP	+ - - -	76	< 975 1006>>	>
ECOTGPRO	+ + + -		<< 587 618 >	<l>
			< 635 664>>	
ECOTGY1	- + - -	77	<< 1563 1587 >	<A
ECOTHR	+ + + -		<< 5397 5414 >	<s>L<s
			< 5422 5453>>	
			<< 5497 5526 >	
ECOTRMF[b]	- + - -	78	<< 31 66 >	<
ECOTRPR	+ + + -		<< 719 744 >	<l>L<
			< 759 788>>	
		31	<< 823 854 >	
ECOTYRB	+ - + -		< 50 79>>	>L[<]
ECOUGP	- - - +	14	< 18 42>>	B>L<A (=UGP1)
		51	<< 87 116 >	
		37	< 1682 1706>>	>S<
		88	<< 1723 1747 >	(=UGP2)
ECOUNG	- - + -	79	< 1237 1262>>	>
ECOUSHA	- - - +		< 1756 1775 >	PU*
ECOUVRB2	+ + + -	10	< 2192 2216>>	B>L<A
		48	<< 2259 2288 >	
ECOUVRD02	+ + + -	11	< 2597 2627>>	>L<
		80	<< 2670 2698 >	
ECOXERB	- - - +	52	< 1595 1624>>	>S<B
		38	<< 1647 1677 >	
ECOXYLE	- + - -	81	<< 2469 2501 >	<A
M22621[d]	- - - +		<< 1436 1544>	S<l>S<l
			< 1470 1501>>	
			<< 1524 1553 >	
STYCYSJIH[d]	- - - +		<< 1754 1785>	<
BIOA[c]	- - - +	82	<< 63 87 >	<
CHLN[c]	- - - +	35	<< 1999 2034 >	<B
CYSA[c,d]	- - - +		<< 3160 3191 >	<
CYSE[c]	- - - +	85	< 1060 1088>>	A>
CYSM[c,d]	- - - +		< 4166 4190>>	B>L<A
			<< 4234 4263>	
FADA[c,d]	- - - +		< 1175 1208>>	>S<l>S<l>S<
			<< 1232 1263 >	
			< 1278 1307>>	
			<< 1329 1360 >	
			< 1375 1404>>	
			<< 1427 1458 >	
GCD[c,d]	- - - +		<< 2571 2595>	<
GLTP[c,d]	- - - +		< 1354 1386>>	>L<s
			<< 1431 1460>	
GYRA[c]	- - - +	84	< 240 269>>	>S<
			<< 289 321 >	
ILVA[c]	- - - +	32	< 6597 6622>>	B>
LIVF[c,d]	- - - +		< 7263 7287>>	B>L<A
			<< 7332 7360 >	
LIVJ[c,d]	- - - +		< 1435 1464 >>	A>L<B
			<< 1509 1533 >	
LPD[c]	- - - +	87	<< 7436 7450 >	<
MALQ[c]	- - - +	40	< 2601 2633>>	>S<
			<< 2658 2684 >	
METF[c]	- - - +	39	< 1072 1099>>	>
M1RNA[c]	- - - +		< 773 801>>	A>L<B
		15	<< 847 871 >	
NAG[c]	- - - +	13	< 3284 3308>>	B>L<A
		50	<< 3352 3381 >	
PHEA[c]	- - - +		< 1797 1820 >	PU*
PURC[c,d]	- - - +		<< 2316 2345>	<
PYRE[c]	- - - +		< 1505 1527 >	PU*
RUV[c]	- - - +		< 2032 2064>>	[<]l>S<
		89	<< 2088 2117 >	
SPEB[c,d]	- - - +		< 1110 1139>>	r1>S<r2>S
			<< 1159 1190 >	(r1=r27c; r2=r26a)
			< 1223 1252>>	
tRNAPHE[c]	- - - +	34	< 97 120>>	>
tRNATYR1[c]	- - - +	33	< 706 731>>	>S<A
		83	<< 753 772 >	

a : The PU are totally included within an open reading freame.
b : The PU is included within the N-formyl methionyl tRNA sequence.
c : These sequences are not referred as in Genbank; the PU position coordinates are numbered as in the following papers. BIOA : 22; CHLN : 23; CYSE : 24; CYSA and CYSM : 25; FADA : 26; GCD : 27; GLTP : 28; GYRA : 29; ILVA : 30; LIVF and LIVJ : 31; LPD : 32; MALQ : 33; METF : 34; M1RNA : 35; NAG : 36; PHEA : 37; PURC : 38; PYRE : 39; RUV : 40; SPEB : 41; tRNAPHE : 42; tRNATYR1 : 43.
d : This sequence was not included in the initial PU cluster collection which was submitted to the "multiple sequence alignment" procedure.
e : The sequence of this region is detailed in figure 8
f : This BIME structure does not include the PU with coordinates <1780 1806>> because this sequence shared very little homology with the PU consensus. Furthermore, the left internal segment of this region (r25 in figure 6) was highly homologous to the B consensus.

Variation Pattern. The homogeneity of a given position i was defined as the highest effective frequency at that position and denoted it H_i (Hi = max EF_i).

'Similarity matrix analysis'

We looked for subsets of similar sequences in larger sets using a procedure of clustering, running on a matrix of similarity scores. First, sequences were ranked in an arbitrary order and numbered accordingly. Secondly, a similarity matrix of these sequences was defined as a n×n array of which the element in row i and column j, denoted sc(i,j), contains the similarity score of sequence i and sequence j. A global score could be computed from that matrix according to the form: GSC = $\Sigma_{i,j}$ $|i-j|$ sc(i,j). Thirdly, since the GSC value decreases when the ranking corresponds to clusters of similar sequences, we wrote a program (ORDONMAT6) designed at minimizing the GSC value.

This program permutes the numbers of the sequences at random, then evaluates the GSC value of the permutated matrix, if the new GSC value is lower than the previous one the program keeps the new numbering, otherwise it reject it. The program stops when a given number of consecutive permutations did not not lead to a lower GSC value. This number of unsuccessful trials is given as a parameter to the program. This procedure does not guarantee to find a numbering of the sequences leading to a global minimum for GSC values, but we ran it using different random changes and found the same sequence clusters.

Hybridization experiments

The hybridization reactions were performed as described (7). The two following *Escherichia coli* strains have been used (7): K12 C600: *thr*1, *leuB*6, *thi*1, *supE*44, *tonA*34, *lacY*1 and CI-1. The CL synthetic probe has the following sequence: 5'-AAAB(T or

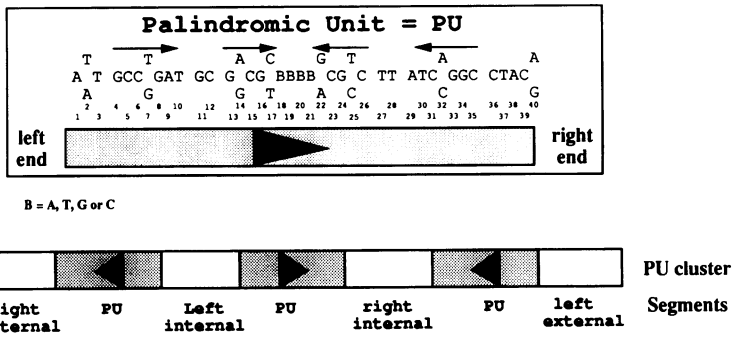

Fig. 1. Structure of a PU cluster. Within the box is shown the PU consensus sequence previously determined in (3). The arrows indicate the positions which exhibit a dyad symmetry. As shown below the consensus, the PU is symbolized by a grey rectangle with an included triangle which indicates the orientation of the PU according to the sequences not included in the symmetry (nucleotides 1–3, 11–12, 27–28 and 36–40). The same PU symbol is used in the other figures. Below the box is indicated a schematic representation of a PU cluster. The external segments are flanked by a single PU. The right external segments which are sequences flanking a the right end of a PU, and the left segments which are sequences flanking the left end of a PU. The internal segments are located between two PU. The right internal segments are sequences located to the right ends of two adjacent PU and the left internal segments are sequences located to the left ends of two adjacent PU.

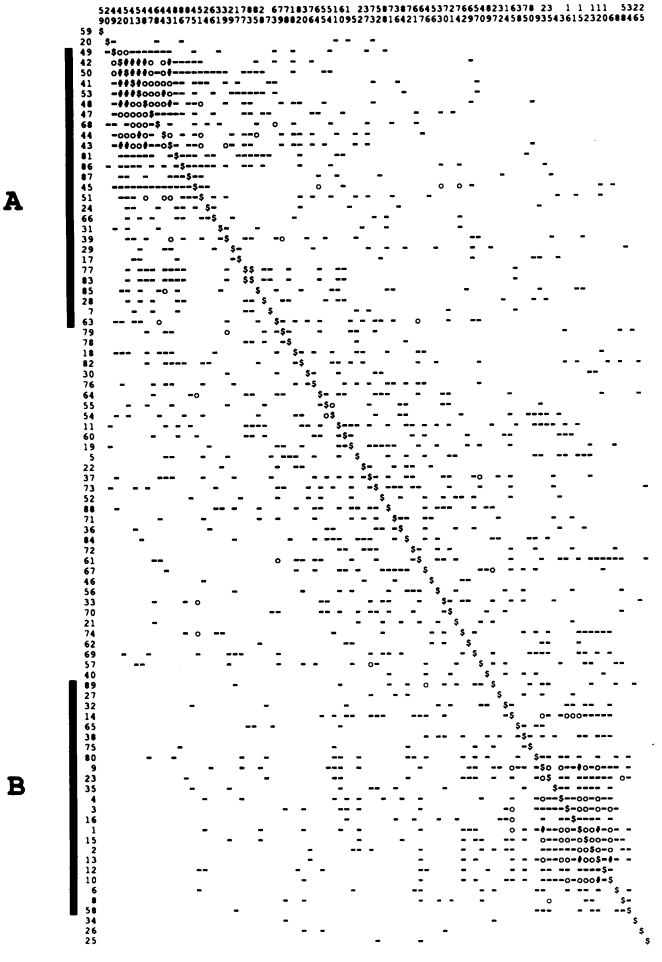

Fig. 2. Clustered similarity matrix for A and B segments. The matrix analysis was used to group the sequences in sets of similarity. The sequences Ordonmat6 program has been fed with the similarity matrix of the left external segments. The program had to try 500 unsuccessful random changes before stopping (see Materials and Methods). Numbers above the matrix and on the right column correspond to sequence numbers. For sake of simplicity, similarity scores less than 50% are not plotted, a – represents scores between 50% and 66%, a o scores between 66% and 82%, a # scores between 83% and 99% and a $ scores equal to 100%. Vertical bars show the two groups of similarity (A and B). Each segment is identified by a number which is referred in table I (column LES).

C)BTGCAAATTCAATA(A or T)ATTGCA(A or G)(A or T)(A or G)(A or T)TCA-3′ (B=A,T,G or C). The sequence of the R and of the C2 probes is given in (7).

RESULTS AND DISCUSSION

We made a compilation of available sequences of *Escherichia coli* PU clusters, i.e; sequence regions containing at least one PU (Materials and Methods). This yielded a collection of 94 clusters, called initial PU cluster set, including 172 PU (table I). In order to analyze the PU surrounding sequences, we divided each PU cluster in segments designated according to their positions with respect to the PU themselves. We then compared the lengths and sequences of segments with similar positions.

Segments located at each end of a PU cluster were called external segments: they are flanked by a single PU. Because PU sequences are oriented, we distinguished the right external segments which are sequences flanking the right end of a PU, and the left external segments which are sequences flanking the left end of a PU cluster (figure 1).

Segments located between two PU were called internal segments. Because of the strict alternance in PU orientations within a cluster (see Introduction), we distinguished the right internal segments as sequences located between the right ends of two adjacent PU and the left internal segments as sequences located between the left ends of two adjacent PU.

Conserved motifs around PU

Two external motifs: A and B. Our PU cluster collection included 51 right external segments and 89 left external segments. In order to detect a possible sequence homogeneity among these segments, we looked for sets of similar sequences by a procedure called 'similarity matrix analysis' (Materials and Methods). We found two sets of similar sequences for the left external segments: groups A and B (fig. 2) which constitute 56% of the total left external segments. By contrast, no such sets were detected for the right external segments (data not shown).

In order to evaluate the sequence homogeneity among the two sets of similar left external segments, we performed a 'multiple

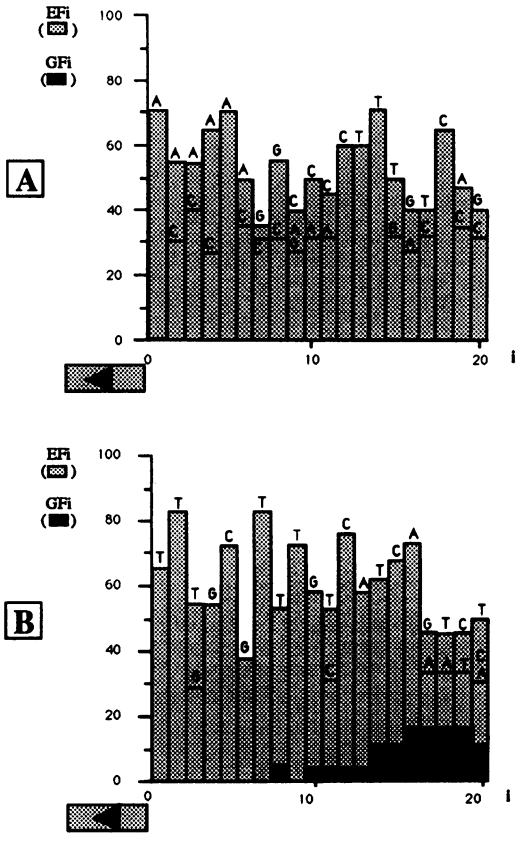

Fig. 3. Sequence Variation Patterns of the left external segments. The Sequence Variation Pattern (SVP), characteristic of a sequence alignment analysis, is a histogram of bases (EFi) and gap frequencies (GFi) at each position i; the definitions of EFi and of GFi are given in Materials and Methods. Below the histogram is shown a schematic representation of the PU (see fig. 1) which is flanking the external segment. Histogram A is the A SVP (28 sequences); histogram B is the B SVP (18 sequences). These histogram have been computed from multiple sequence alignment with a gap penalty of 2.0 +3.0l (see Materials and Methods).

sequence alignment' with each of them. In both cases an homogeneity, averaging 55% for A and 60% for B (fig. 3A and 3B) was found along most of the analyzed regions.

This led us to define two conserved motifs (A and B, fig.2) corresponding to the consensus sequence of each set.

Five internal motifs: S, L, s, l and r. Our PU cluster collection contained 51 right internal segments and 28 left internal segments.

The size distribution of the right internal segments presents two modes (fig. 4A). About 50% of the segments have a size ranging from 32 to 34 nucleotides (centered at 33); they were called L (for Long sequences). The L set exhibited a high sequence homogeneity, averaging 80% (fig. 5B). The other segments have a size comprised between 11 and 14 nucleotides (centered at 12); they were called S (for Short sequences). The S sequences presented a high level of size homogeneity (see their sharp mode of distibution in figure 4A), and a limited sequence homogeneity, averaging 50% (figure 5A).

The size distribution of the left internal set presents three modes (fig. 4B). 8 out of the 28 left internal segments have a size of 8 or 9 nucleotides; they were called l (for *l*ong sequences). The l set exhibited a significant sequence homogeneity, averaging 68% (fig. 5C).11 out of the 28 segments are composed of a single nucleotide: cytosine in 60% of the cases; they were called s (for short sequences). A third mode, much more heterogeneous in size, can be defined with the 5 larger segments; they have a size ranging from 18 to 31 nucleotides; they were called r. The r set did not reveal any sequence homogeneity (data not shown).

By eye inspection, we found some homologies between the r sequences and between the A and the B external consensus sequences (fig. 6). Each r sequence could be divided in two parts; the right part exhibits homologies with the part of the A consensus which is just flanked by a PU; the left part exhibits homologies with the part of the B consensus which is just flanked by a PU (fig. 6).

In conclusion, the sorting of internal segments in five sets based on size and sequence led us to define five motifs. Four of them (L, S, l, and s) have sharply defined sizes; among these, three

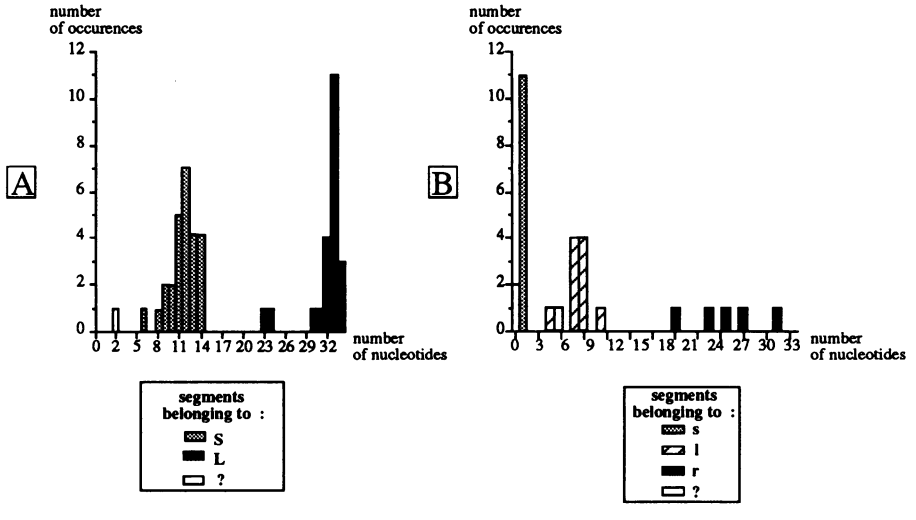

Fig. 4. Size distribution of the internal segments. The diagrams indicate the number of sequences present in each set (see text) as a function of their size. Diagram A: size distribution of the right internal segments; diagram B: size distribution of the left internal segments. Below the diagrams are boxed the symbols used to show the distribution of the the different segments described in the text.

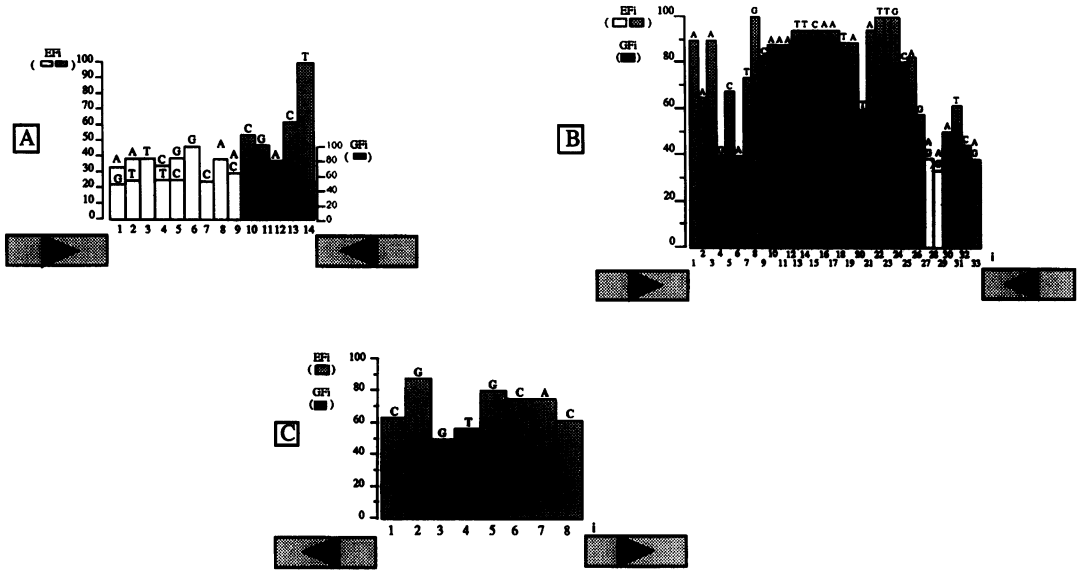

Fig. 5. Sequence Variation Patterns of the internal segments. See fig. 3. Below the histogram is shown a schematic representation of the PU (see fig. 1) which are flanking the internal segment. Histogram A is the S SVP (20 sequences): only 5 positions (10–14, colored in grey) out of 14 exhibit a significant homogeneity; histogram B is the L SVP (18 sequences): only 2 positions (28–29, in white) out of 33 do not exhibit a significant homogeneity; histogram C is the l SVP (8 sequences). The gap penalty used in the multiple alignment were 2 + 1.5 l and 3 + 2 l respectively for S and L segments and for s and l segments.

```
A CONSENSUS

     AAAAAAGGCCCCTTTGTCAGAAATTA
      CCC CCCAAA   GAC CCC  A T
           G
        T↓ A↓
     ATGTGCTGCCCC--GATCC                    r19
           G↓
     AAAATTTCAGGCTTTATGAGT-ATTT             r23
     AAT---GGCACATTTGTTACCTTGTGCGC          r26a
     AAT---GGCACGTTTT--ACCC-GTGCGCATCG      r27c
     GAACAGACAAACAGTTTCAAACGCTAA            r27a
     AAACAG-CAAACAATCCAAAACGCCGC            r26b
        T↓
     AAGCAATAAGACATGGTTAGCTTTATATTTG         r31

B CONSENSUS

          GAGACTGATGACAAACGCAAA
          CTATT     G       C
            G
     ATGTGCTTGCCACCGATCC                    r19
          ATTTCAGGCGTTTATGAGTATTT            r23
     AATGGCACATTTGTTACCTTGTGCGC             r26a
     AATGGCACGTTTT--ACC-CGTGCGCATCG         r27c
       GAACAGACAAACAGTTTCAAACGCTAA          r27a
       AAACAG-CAAACAATCCAAAACGCCGC          r26b
                TGGTT↓
          AAGCAAT-AAGTACAAGCTTTATATTTG       r31
```

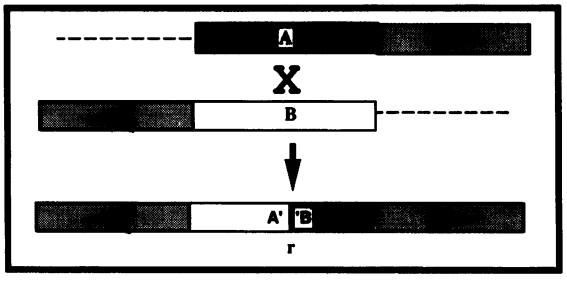

Fig. 6. The r segment can be obtained by recombination between the A and B segments. The r segments have been aligned with the A consensus (upper part of the figure) and with the B consensus (middle part of the figure). The underlined positions are identical to the corresponding consensus. The small vertical arrows indicate an insertion on the left of the nucleotide toward the arrow is pointed. In the lower part of the fig. is boxed a model for the formation of the r segment by a recombination between a A and a B segment (A' 'B).

have a strict consensus sequence (L, S, and l) while the consensus sequence of the fourth is loose. In the case of the fifth motif (r), length and sequence are variable, but r can be considered at least formally as a combination between the A and B motifs. This suggests the hypothesis that they are indeed generated by recombination between A and B motifs (lower part of fig.6).

Species-specific hybridization of an L motif probe

We designed a synthetic probe according to the L motif consensus (named CL; Materials and Methods). Many fragments of the total *Escherichia coli* DNA cleaved with *HinfI*, strongly reacted with the CL probe (fig.7A). This pattern was similar to that obtained with a PU specific probe (C2 probe) both in intensity and in number of fragments (fig.7A) (7). We used, as a negative control, a synthetic probe with a random primary sequence (R probe)(7). Since this probe hybridized poorly (fig. 7A), we concluded that we detected a specific cross-hybridization between CL and *Escherichia coli* genomic DNA. This result confirms that the L motif corresponds to a repetitive DNA with a high level of sequence homogeneity along the genome.

In order to see whether the L motif was present in other bacteria, we performed a Southern experiment with the genomic DNA of ten different bacteria (fig. 7B). Only *Escherichia coli* and *Shigella sonnei* genomic DNA exhibited a significant cross-hybridization with the CL probe (compare lanes 1 and 10 with lanes 2 to 9 in fig. 7B). Even *Salmonella enteritica* serotype Typhimurium, which is closely related to *Escherichia coli*, did not cross-react with the CL probe (lane 9, fig. 7B). These results clearly showed that the presence of L sequences, as reflected by multiple cross-hybridization with the CL probe, is bacterial species-specific. It is interesting to note that PU species-specificity correlates well with that of the L sequences: *Escherichia coli* and *Shigella sonnei* possess the same type of PU sequence, different from the type present in other *Enterobacteriaceae* (Introduction

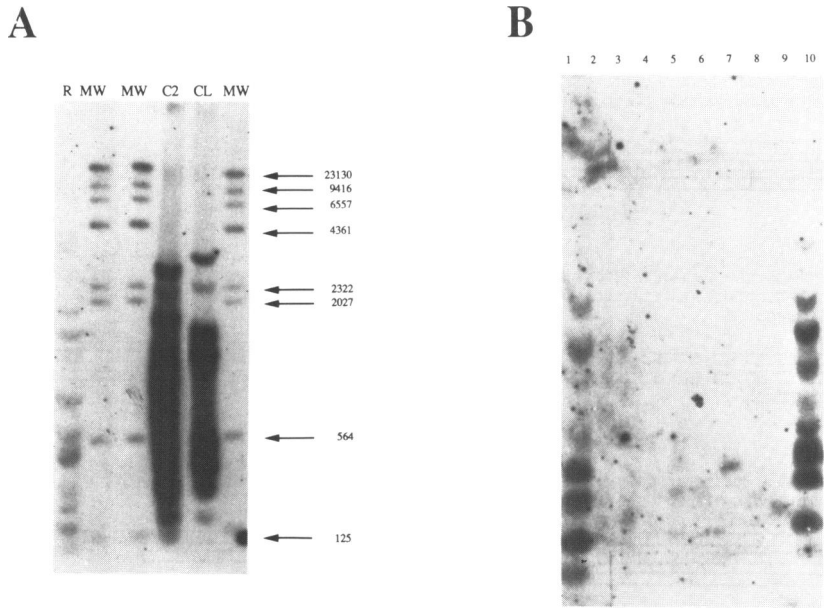

Fig. 7. Southern hybridization of CL probe with genomic DNA of enterobacteria. The hybridization experiments were performed as described in Materials and Methods. Fig. 7A. The total genomic DNA of *Escherichia coli* C600 (500 fmoles) was hybridized with either R or C2 or CL as indicated above each lane. The lanes called MW (for Molecular Weight) contain the *Hind*III digest of lambda DNA; the size of the fragments is indicated on the figure. Fig. 7B. Each lane contain 100 fmoles of total genomic DNA of the following bacteria. Lane 1: *Escherichia coli* CI-1; lane 2: *Yersinia pseudotuberculosis* P105; lane 3: *Cedecea davisae* CIP80.34; lane 4: *Klebsiella pneumoniae* K2; lane 5: *Enterobacter cloacae* (a gift of the 'Centre Hospitalier de Villeneuve St Georges, France'); lane 6: *Enterobacter aerogenes* ATCC 3048; lane 7: *Citrobacter freundii* CDC 22−76; lane 8: *Levinea malonatica* CDC 1066-71; lane 9: *Salmonella enteritica* serotype Typhimurium LT2; lane 10: *Shigella sonnei* S60-80. When no collection is indicated, the strains were from the Unité des Entérobactéries, Institut Pasteur, Paris, France. CDC, Center for Disease Control, Atlanta, Georgia, CIP: Collection Institut Pasteur.

and 7). In this respect, it would be interesting to know whether other types of L sequences existed in other bacteria.

PU clusters are a mosaic combination of 8 sequence motifs: the BIME

The above results show that PU clusters are a combination of a limited number of sequence motifs. The total number of motifs detected so far is eight: the PU sequence itself and 7 PU surrounding motifs (A, B, S, L, s, l and r). All these motifs, except r, are homogeneous in size and/or in sequence (see above). The r motif appears to be formed by one part of a A sequence and by one part of a B sequence.

There are two remarkable properties of these motif combinations. Firstly, all the PU clusters, from the initial collection, are a mosaic combination of these and only these eight motifs. Furthermore, after the identification of these motifs, we found 19 new PU clusters in subsequent publications (earmarked d in Table I); all were combinations of these eight motifs. Secondly, we searched in an *Escherichia coli* sequence data base for the longer of the motifs located outside a PU (A, B and L): they were always found associated with at least one PU, showing that they are characteristic of PU clusters (unpublished data).

Because PU clusters appear to be mosaic combinations of the same eight sequence motifs, we propose to call them BIME for Bacterial Interspersed Mosaic Element. Our present total collection of BIME amounts to 113 with 236 PU sequences (table I), corresponding to 20% of the total *Escherichia coli* genomic sequence; it can be extrapolated that this genome contains about 500 BIME. If we estimate that the *Escherichia coli* genome is composed of 3000 genes, about one extragenic region in six contains a BIME.

Some BIME are a direct repetition of a PU doublet

The mode of combination of the eight motifs generates a great diversity of BIME structures (Table I). The 113 BIME include from a single motif (one PU) to complex structures, such as in ECOARAABD (symbolyzed as S<s>S<s>S<s>S<s where the PU sequences are indicated by a '<' or a '>' according to their orientation; see table I). Remarkably, the diversity in BIME motif combinations can be accounted for by a simple rule: within a BIME, all the occurrences of the same type of internal segment (left or right) consist in the same motif; in other words, all the right internal segments are either S or L; all the left internal segments are either s or l or r. Furthermore, BIME appear to be composed by an array of direct repetitions since the homogeneity of any motif is higher within a BIME than between BIME (some examples are shown in figure 8).

From further inspection of the different types of direct repetions present in BIME, we inferred that each BIME can be described as an array of a PU doublet repeats. Some examples are shown in figure 8; as illustrated by the alignment presented, they are all composed by a repetition of a PU doublet.

These observations could be explained by a localized amplification of a PU doublet. A shematic representation of this hypothesis is presented in fig.9. If all the BIME were initially formed by a repetition of a PU doublet, BIME with an odd number of PU (about 50% of the known BIME) could be explained, for example, by deletion of an odd number of PU from the initial structure.

Interestingly, the localized amplification events appear to operate sometimes on sequences just flanking a BIME. Indeed, some BIME are located just at the boundary of a DNA region which exhibits direct repetitions. This is the case in the BIME

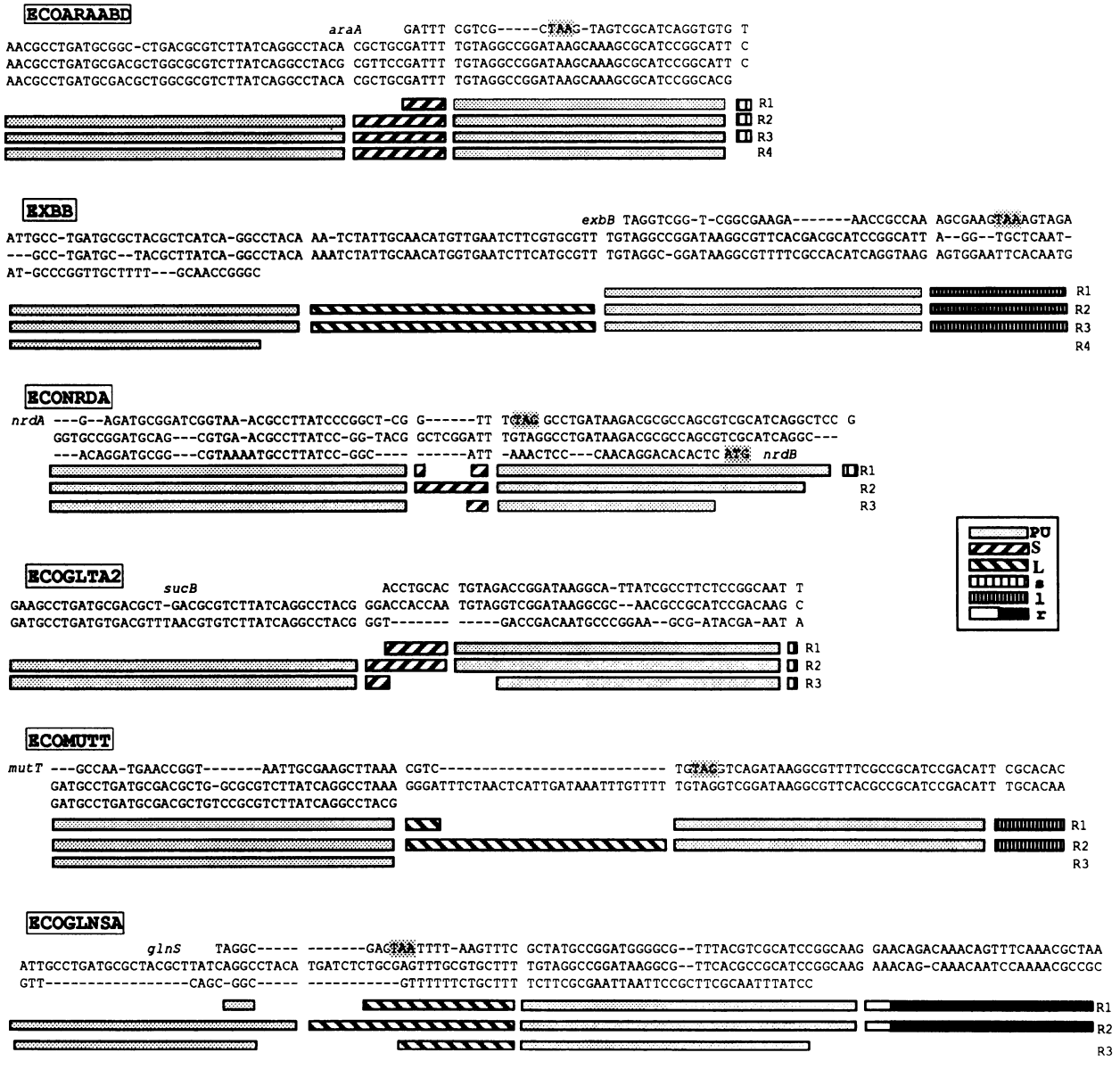

Fig. 8. Direct repetitions in some BIME. Each sequence is decomposed in different BIME segments. The name of the BIME (Table I) is boxed. The start and stop codons of the flanking open reading frames are overlined in grey. The sequence has been rearranged in order to show the internal repetitions (called R1, R2 R3 and R4). Below the sequence is a schematic representation of the different segments. The legend of the symbols used for each segment is boxed on the right of the figure.

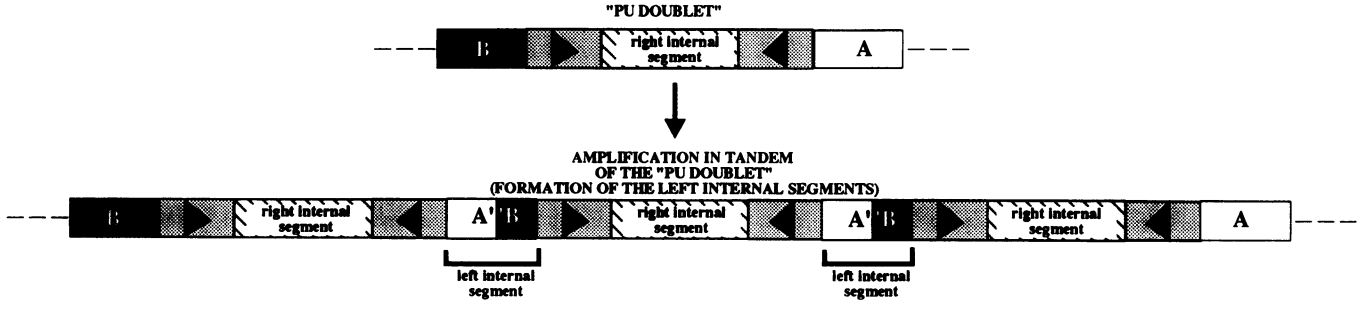

Fig. 9. Hypothesis for BIME associated localized amplification. The hypothesis predicts that a PU doublet (see text), schematically represented at the top of the figure, can undergo multiple rounds of amplification, leading to amplified BIME structures, as shown at the bottom of the figure. During the amplification process, the joint segments between the PU doublet repeats constitute new sequences which are described as left internal segments in the text. At least one of the motifs present in the left internal segments (the r motif) can simply be explained by a recombination between a A and a B sequence (as shown in the fig.6).

flanking region of ECOTGPRO, of tRNATYR1 and of M1RNA (table I). The presence of this type of directly repeated DNA is quite exceptional in *Escherichia coli*. This association suggests that the presence of a PU doublet is able to stimulate the formation of amplified DNA not only at BIME but also in its immediate vicinity (for a review on amplification see 19).

It is interesting that one of the two proteins known to interact specifically with a BIME motif, DNA pol I (20), is believed to be involved in some amplification processes (21). Thus, a specific BIME-DNA pol I interaction could favour the formation of localized amplified sequences at or near BIME.

BIME functions

The abundance of PU suggested that they may serve one (or several) physiologically important function(s). Their sequence homogeneity could indicate at least a common feature in the role of all PU. There are numerous processes where PU were shown to play a part. This include transcription termination (8), mRNA stabilization (9, 10), control of translation (11), genomic rearrangements (2,12) and nucleoid folding (3, 13, 14).

Since PU sequences are part of BIME, a larger genetic element which includes other sequence motifs, the question is raised of the particular role of each motif in each of these functions. We suggest that BIME could belong to several functional classes depending on the nature of the motifs. For example, as the PU portion of BIME has been shown to specifically interact with DNA pol I and gyrase (see 'Introduction'), some of these classes could lead to the formation of different nucleoprotein complexes. The newly defined motifs have not all been tested for their capacity to bind proteins. It will be interesting to see if the L motif, which is highly conserved, has affinity for a new protein.

ACKNOWLEDGMENTS

This work was supported by grants from the Fondation pour la Recherche Médicale, the Ligue Nationale contre le Cancer, and the Association pour la Recherche sur le Cancer.

REFERENCES

1. Higgins, C.F., Ames, G.F.-L., Barnes, W.M., Clément, J.M. & Hofnung, M. (1982). *Nature*, **298**, 760–762.
2. Gilson, E., Clément, J.M., Brutlag, D. & Hofnung, M. (1984).*EMBO J.* **3**, 1417–1421.
3. Gilson, E., Clément, J.M., Perrin, D. & Hofnung, M. (1987).*Trends Genet.* **3**, 226–230.
4. Merino, E. & Bolivar, F. (1989). *Mol. Microbiol.* 3, 839–841.
5. Stern, J.M., Ferro-Luzzi Ames, G., Smith, N.H., Robinson, E.C., and Higgins, C.F. (1984). *Cell*, 37, 1015–1026.
6. Higgins, C.F., Mc Laren, R.S. & Newbury, S. (1988). *Gene* **72**, 3–14.
7. Gilson, E., Bachellier, S., Perrin, S., Perrin, D., Grimont, P.A.D., Grimont, F., and Hofnung, M. (1990), *Res. Microbiol.***141**, 1103–1116.
8. Gilson, E., Rousset, J.P., Clément, J.M. & Hofnung, M. (1986*a*). *Ann. Inst. Pasteur/Microbiologie.* **137B**, 259–270.
9. Newbury, S.F., Smith, N.H., Robinson, E.C., Hiles, I.D. & Higgins, C.F. (1987*a*). *Cell*, **48**, 297–310.
10. Newbury, S.F., Smith, N.H. & Higgins, C.F. (1987*b*). *Cell*, **51**, 1131–1143.
11. Stern, M.J., Prossnitz, E. & Ames, G. F.-L. (1988). *Mol. Microbiol.* **2**, 141–152.
12. Shyamala, V., Schneider, E. & Ames, G. F.-L. (1990). *EMBO J.*, **9**, 939–946.
13. Gilson, E., Perrin, D., Clément, J.M., Szmelcman, S., Dassa, E. & Hofnung, M. (1986*b*). *FEBS Lett.* **206**, 323–328.
14. Yang, Y. & Ames, G.F.-L. (1988). *Proc. Natl. Acad. Sci. USA* , **85**, 8850–8854.
15. Claverie, J.M. & Sauvaget, I. (1985). *Cabios*, **1**, 95–104.
16. Lipman, D.J. & Pearson, W.R. (1985). *Science*, **227**, 1435–1440.
17. Gribskov M, Lüthy R. & Eisenberg D. (1990). In Doolittle R. F. (ed.), Methods in Enzymology–Molecular evolution: Computer analysis of protein and nucleic acid sequences., Vol. **183**, pp. 146–159
18. Gotoh, O. (1982). *J. Mol. Biol.*, 162, 705–708.
19. Stark, G.R. & Wahl, G.M. (1984) *Ann. Rev. Biochem.*, **53**, 447–491.
20. Gilson, E., Perrin, D. & Hofnung, M. (1990). *Nucleic Acids Res.*, **18**, 3941–3952.
21. Kornberg, A. (1980). *DNA Replication.* W. H. Freeman and Co, San Fransisco, pp 101–166.
22. Otsuka, A.J., Buoncristiani, M.R., Howard, P.K., Flamm, J., Johnson, C., Yamamoto, R., Uchida, K., Cook, C., Ruppert, J. & Matsuzaki, J. (1988). *J. Biol. Chem.* **263**, 19577–19585.
23. Nohno, T., Kasai, Y. & Saito, T. (1988). *J. Bacteriol.* **170**, 4097–4102.
24. Denk, D. & BÖck, A. (1987). *J. Gen. Microbiol.* 133, 515–525.
25. Sirko, A., Hryniewicz, M., Hulanicka, D. & BÖck, A. (1990). *J. Bacteriol.* 172, 3351–3357.
26. Yang, S.Y., Yang, X.Y.H., Healy-Louie, G., Schulz, H. & Elzinga, M. (1990). *J. Biol. Chem.*, **265**, 10424–10429.
27. Cleton-Jansen, A.M., Goosen, N., Fayet, O., & van de Putte, P. (1990). *J. Bacteriol.* **172**, 6308–6315.
28. Wallace, B., Yang Y.J., Hong, J. & Lum, D. (1990). *J. Bacteriol.* **172**, 3214–3220.
29. Hussain, K., Elliot, E.J. & salmond, G.P.C. (1987). *Mol. Microbiol.* 1, 259–273.
30. Wek, R.C. & Hartfield, G.W (1986). *J. Biol. Chem.*, **261**, 2441–2450.
31. Adams, M.D., Wagner, L.M., Graddis, T.J., Landick, R., Antonucci, T.K., Gibson, A.L. & Oxender, D.L. (1990). *J. Biol. Chem.* **265**, 11436–11443.
32. Stephens, P.E., Lewis, H.M., Darlison, M.G. & Guest, J.R. (1983). *Eur. J. of Biochem.*, 135, 519–527.
33. Pugsley, A.P. & Dubreuil, C. (1988). *Mol. Microbiol.* **2**, 473–479.
34. Saint-Girons, I, Duchange, N., Zakin, M., Park, I., Margarita, D., Ferrara, P. & Cohen, G. (1983). *Nuc. Acids Res.*, **11**, 6723–6732.
35. Reed, R. E. & Altman, S. (1983). *Proc. Natl. Acad. Sci. USA*, **80**, 5359–5363.
36. Plumbridge, J.A. (1989). *Mol. Microbiol.* **3**, 505–515.
37. Hudson, G.S. & Davidson, B.E. (1984). *J. Mol. Biol.* **180**, 1023–1051.
38. Tiedeman, A.A., DeMarini, D.J., Parker, J., & Smith, J.M. (1990). *J. Bacteriol.* **172**, 6035–6041.
39. Lundberg, L.G., Thoresson, H.O., KarlstrÖm, O.H. & Nyman P.O. (1983). *EMBO J.*, **2**, 967–971.
40. Shinagawa, H., Makino, K., Amemura, M., Kimura, S., Iwasaki, H. & Nakata, A. (1988). *J. bacteriol.* **170**, 4322–4329.
41. Szumanski, M.B.W. & Boyle, S.M. (1990). *J. Bacteriol.* **172**, 538–547.
42. Schwartz, I, Klotsky, R.-A., Elseviers, D., Gallagher, P.J., Krauskopf, M., Siddiqui, M.A.Q., Wong, J.F.H. & Roe, B.A. (1983). *Nuc. Acids Res.*, **13**, 4379–4389.
43. Mc Corkle, G.M. & Altman, S. (1982) *J. Mol. Biol.* **155**, 83–103.