Characterization of an Unusually Conserved AluI Highly Reiterated DNA Sequence Family From the Honeybee, Apis mellifera

Sophie Tarès,* Jean-Marie Cornuet† and Pierre Abad**1

*Laboratoire de Biologie des Invertébrés, Institut National de la Recherche Agronomique, 06606 Antibes Cédex, France, and
†Laboratoire de Neurobiologie Comparée des Invertébrés, Institut National de la Recherche Agronomique-Centre National de la Recherche Scientifique, 91440 Bures-sur-Yvette, France

Manuscript received September 15, 1992
Accepted for publication April 14, 1993

ABSTRACT

An AluI family of highly reiterated nontranscribed sequences has been found in the genome of the honeybee Apis mellifera. This repeated sequence is shown to be present at approximately 23,000 copies per haploid genome constituting about 2% of the total genomic DNA. The nucleotide sequence of 10 monomers was determined. The consensus sequence is 176 nucleotides long and has an A + T content of 58%. There are clusters of both direct and inverted repeats. Internal subrepeating units ranging from 11 to 17 nucleotides are observed, suggesting that it could have evolved from a shorter sequence. DNA sequence data reveal that this repeat class is unusually homogeneous compared to the other class of invertebrate highly reiterated DNA sequences. The average pairwise sequence divergence between the repeats is 2.5%. In spite of this unusual homogeneity, divergence has been found in the repeated sequence hybridization ladder between four different honeybee subspecies. Therefore, the AluI highly reiterated sequences provide a new probe for fingerprinting in A. m. mellifera.

HIGHLY repetitive DNAs have now been characterized in a number of plant and animal species (SKINNER 1977; JELINEK and SCHMID 1982; BERTIDZE 1986). They constitute very simple closely related sequences that are repeated thousands of times in tandem arrays. Their abundance can vary from less than 1% to more than 66% of the genome (SKINNER 1977), and their complexity ranges from 2 to 2,000 bp (MIKLOS 1982; MIKLOS and GILL 1982; FOWLER et al. 1985). In general, these sequences are concentrated in centromeric and telomeric regions of a chromosome and constitute the principal component of heterochromatin (MIKLOS 1985; DAVIS and WYATT 1989). They are not transcribed. Due to their ability to exhibit variation within and between subspecies, they offer a good prospect for population or individual identification (FOREJT 1973; MACKAY, BOBROW and COOKE 1978). In spite of great sequence divergence between closely related species, great similarity has often been observed within the repeats of one species (HSIEH and BRUTLAG 1979). Therefore, study of highly repetitive DNAs may be a way to analyze mechanisms of genome dynamics not influenced by sequence-specific selection pressure.

Apis mellifera represents both an insect of high agronomical value (pollination of many important crops) and an important biological model due to its learning behavior. However, few studies at the DNA level have been reported in this species. Most of them have analyzed mitochondrial DNA (mtDNA) in order to discriminate subspecies (SMITH 1988; SMITH and BROWN 1988, 1990; HALL and SMITH 1991), to establish the evolutionary history of this species (SMITH 1991a,b; CORNUET and GARNERY 1991; GARNERY, CORNUET and SOLIGNAC 1992), to infer phylogenetic links between species of the same genus (GARNERY et al. 1991), and to study a specific intergenic sequence which might play a part in the replication of the molecule (CORNUET, GARNERY and SOLIGNAC 1991). mtDNA studies have also contributed to a better understanding of the “africanization” process of American populations (HALL and MURALIDHARAN 1989; SMITH, TAYLOR and BROWN 1989). Partial sequences of mtDNA have been published by VLASKACK, BURGSWAIGER and KREIL (1987), CROZIER, CROZIER and MACKINLEY (1989) and now the complete sequence is available (CROZIER and CROZIER 1993).

At the nuclear DNA level, studies concern genome complexity (JORDAN and BROSEMER 1974), patterns of DNA sequence arrangement which are similar to those of Drosophila melanogaster, the genome organization, which is composed of 90% of single copy sequences (CRAIN, DAVIDSON and BRITTEN 1976), and characterization of unique sequences such as homeobox-containing genes (WALLDORF, FLEIG and GERRING 1989). However, studies concerning the organization of repetitive DNA are lacking in A. mellifera, perhaps because less than 10% of the genomic DNA

1 To whom reprint requests should be addressed.

Genetics 134: 1195–1204 (August, 1993)
is repetitive (CRAIN, DAVIDSON and BRITTEN 1976).

Here we report the cloning and characterization of a member of a highly repeated DNA family from the honeybee *A. mellifera*. We have analyzed its abundance, its sequence organization, primary structure and transcriptional activity. This repetitive sequence is shown to represent 2% of the *A. mellifera* genome and is not transcribed in any tissue. It exhibits a high homogeneity, the 10 sequenced monomers showing an average of 1.4% divergence from the consensus sequence. However, in the hybridization ladder obtained with this repeat, DNA restriction fragment length polymorphisms (RFLPs) have been identified between the four *A. mellifera* subspecies tested and within the *A. m. mellera* race suggesting that it could be used as a genetic marker for *A. mellifera* population genetic studies.

**MATERIALS AND METHODS**

**Materials**: Honeybees (*A. mellifera*) subspecies were collected from colonies of *A. m. ligustica*, *A. m. caucasia*, *A. m. mellifera* and *A. m. scutellata*. Lineages of *A. m. ligustica* and *A. m. caucasia* were maintained by artificial insemination while *A. m. mellifera* came from local colonies of southeast France. DNA samples from each of these subspecies were represented by workers from one colony per subspecies, except for *A. m. mellifera* where two workers of the same *A. m. mellifera* colony (no. 172), one worker of *A. m. mellifera* colony (no. 146), one worker of *A. m. mellifera* colony (no. 280) have been tested. *A. m. scutellata* workers came from a colony of South Africa. In addition, workers from one colony of two other *Apis* species from Thailand (*A. cerana* and *A. dorsata*) have been isolated. Radionucleotides were purchased from Amersham and restriction endonucleases from Boehringer Mannheim.

**DNA analysis**: Methods for DNA preparation from adult workers, restriction endonuclease digestions, electrophoresis, transfer to nitrocellulose or nylon membranes, prehybridization and hybridization conditions have been described previously (ABAD et al. 1989; BUCHETON et al. 1984; MANIATIS, FRITSCH and SAMBROOK 1982). Hybridizations were performed at 65°C. Conditions for high stringency consisted of washing at 65°C in a 0.1× SSC-0.1% sodium dodecyl sulfate (SDS) final solution.

**Isolation and cloning of highly repeat sequences**: Honeybee *A. mellifera* genomic *Alu* I fragments less than ~200 bp in length were isolated from a 1% agarose gel and ligated into Smal-digested pUC 19 (Appligen). One recombinant plasmid (Ap1), containing a 176-bp insert, was used as probe to isolate other repeats and to hybridize total genomic DNA in Southern blots.

**RNA preparations**: Total and poly(A)^+^ RNAs from adult *A. m. mellifera* were isolated as described earlier (MANIATIS, FRITSCH and SAMBROOK 1982). Northern hybridization was conducted under the same conditions as for DNA.

**Nucleotide sequencing**: The nucleotide sequences of 10 cloned highly repeated sequences monomers were determined by the dideoxynucleotide chain-termination method (SANGER, NICKLEN and COULSON 1977) using double-stranded plasmid templates and Sequenase kit (U.S. Biochemical Corp.).

**Analysis of RFLP data**: Autoradiograms were used to determine variations in RFLP patterns of honeybees from different colonies of *A. m. mellifera* and different subspecies of *A. mellifera*. Similarity of coefficients for all pairwise combinations of RFLP groups were calculated using the formula

$$ F = 2N_x/N_x + N_y $$

where $N_x$ = number of bands in genotype $x$ and $N_y$ = number of bands shared by genotypes $x$ and $y$ (NITZ and LI 1979). $F$ ranged from 0 (no shared bands) to 1 (all bands common to both genotypes). The distances between the different patterns seen in Figure 7B were deduced from $F$ values according to the formula

$$ d = 1 - F $$

**Nucleotide distances and phylogenetic analysis**: Nucleotide distances between sequences have been computed according to KIMURA (1980), with deletions counted as transversions. The neighbor-joining algorithm (SAITOU and NISHI 1987) has been used to build phylogenetic trees of repeat sequences. Since no outgroup was available, trees have been artificially rooted by folding the network at the midpoint of the largest patristic distance between repeats.

**RESULTS**

**Isolation of a honeybee *Alu*I highly repeat DNA sequence family**: Digestions of genomic *A. m. mellifera* DNA with various restriction endonucleases could not reveal the presence of strong bands in agarose gel electrophoresis after ethidium bromide staining. However, the corresponding Southern blot, hybridized with nick-translated total genomic *A. m. mellifera* DNA, exhibits a strong band of approximately 180 bp and several bands of multimeric lengths in the *Alu*I digestion (Figure 1).

The *Alu*I monomer band was purified from agarose gel and subcloned into pUC 19 vector. Ten clones were isolated for further analysis. They have an average length of 176 bp. One of these clones, Ap1 was used to hybridize genomic Southern blots (Figure 2). The probe lights up a typical ladder pattern of the repeated sequences arranged in tandem arrays. The *Alu*I pattern shows that part of the repeated DNA remained as oligomers (up to tetramer) of the 176-bp repeat unit, indicating tandem repetition of the monomer due to loss of the *Alu*I site in some of the units. By contrast, digestions with other enzymes produce a reverse ladder with darker bands caused by larger DNA fragments, restriction sites having appeared by mutation in some repeats.

Furthermore, in the Southern analysis of genomic DNA digested with *Alu*I, one can notice the presence of strong hybridization bands outside the repeat ladder of the 176-bp monomer. These bands are approximately 600 and 820 bp long, respectively. The hybridizations occur even with high final washing stringency (0.1× SSC-0.1% SDS). This result indicates that these DNA bands are more than 80% identical to the *Alu*I highly repeat DNA class, but do not produce oligomers after *Alu*I digestion.
Quantification and copy number of the \textit{AluI} repeat in the genome: Increasing amounts of cloned monomer were dot blotted onto nitrocellulose membrane and hybridized with the satellite monomer Ap1 (Figure 3). PUC18 was used as a background control and genomic \textit{D. melanogaster} and \textit{Caenorhabditis elegans} DNAs were as negative controls.

The \textit{AluI} highly reiterated sequence family appears to make up 2% of the total genomic DNA as calculated from values obtained by scintillation measurements. Taking into account the length of the satellite and a \(C\) value of 209,000 kb for \textit{A. mellifera} genome (JORDAN and BROSEMER 1974; CRAIN, DAVIDSON and BRITTEN 1976), this fraction corresponds to approximately 23,000 copies per haploid genome.

Primary structure of the satellite repeats: The nucleotide sequence from each of ten cloned monomers was determined in order to elucidate the internal organization of the repeats (Figure 4). The lengths of these sequences were found to be quite homogeneous, varying between 173 and 180 bp, which closely agrees with the estimate based on electrophoretic mobility of the restriction fragments. An unambiguous consensus sequence of 176 bp long was derived from the complete data and is listed at the top of the Figure 4. The identity of this consensus sequence to that of clone Ap3 is fully consistent with the basal position of this clone in the phylogenetic tree shown in Figure 8. The DNA has an A+T content of 58%. The 10 repeats differ by single base substitutions or deletions and by a small insertion. We have also analyzed our sequences with respect to distribution of several restriction sites used on Southern blot (Figure 2) which reveal a typical ladder pattern. None of them is present in the 10 sequenced monomers, except for \textit{AluI} and \textit{DraI}. This
<table>
<thead>
<tr>
<th>Monomer length</th>
<th>Sequence and restriction enzyme recognition sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus sequence 176 bp</td>
<td>1 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170</td>
</tr>
<tr>
<td>Ap1 176 bp</td>
<td></td>
</tr>
<tr>
<td>Ap2 176 bp</td>
<td></td>
</tr>
<tr>
<td>Ap3 176 bp</td>
<td></td>
</tr>
<tr>
<td>Ap4 176 bp</td>
<td></td>
</tr>
<tr>
<td>Ap5 175 bp</td>
<td></td>
</tr>
<tr>
<td>Ap6 180 bp</td>
<td></td>
</tr>
<tr>
<td>Ap7 175 bp</td>
<td></td>
</tr>
<tr>
<td>Ap8 173 bp</td>
<td></td>
</tr>
<tr>
<td>Ap9 176 bp</td>
<td></td>
</tr>
<tr>
<td>Ap10 176 bp</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.—Consensus sequence (on top) and sequences of 10 individual clones of AluI monomers. The reference sequence represents the consensus sequence deduced from all sequenced DNA fragments. Base substitutions relative to the reference are indicated by the respective bases, deletions by (−) and insertion by bracket. The restriction enzyme recognition sites are underlined and common restriction sites in all monomers, AluI and DraI, are shaded.
In this report, we have identified and described for the first time the molecular features of a highly repeated DNA sequence from honeybee. We have cloned monomers resulting from digests of genomic A. m. mellifera DNA with the restriction enzyme AluI. The ladder pattern was also observed with other restriction enzymes. Ten Alu fragment monomers were sequenced. Due to random selection, they are expected to be a representative sample for the most abundant sequence of the fragment type. As our results show, each clone corresponds to a different variant of the same related sequence. Sequence analysis of the different monomers shows a great homogeneity between members of this highly reiterated DNA class. It should be pointed out that sequence analysis of monomers indicates the presence of internal recognition sites only for DraI but not for any other enzyme which give the typical ladder pattern on Southern blot (Figure 2). This could be explained by the fact that each of these last enzymes has a recognition site in about less than 5% of the total satellite integrations corresponding to the monomer API. Figure 7A reveals the presence of the AluI tandem repeats family in the sibling species to A. mellifera, A. cerana and its absence in A. dorsata.

**DISCUSSION**

To analyze the distribution of this tandemly repeat sequence in the Apis genus, we have also hybridized the genomic DNA of two other species of Apis with one of the monomers (API). Figure 7A reveals the presence of the AluI tandem repeats family in the sibling species to A. mellifera, A. cerana and its absence in A. dorsata.

**Polymorphism between geographical groups of honeybees:** The Southern blot hybridization pattern obtained by AluI digestion of genomic DNA of two workers from three different A. m. mellifera colonies and three other A. mellifera subspecies indicates that this repetitive sequence is able to discriminate between them. In the extended Alu ladder pattern, the numbers and positions of bands indicate polymorphism between individuals from different colonies of A. m. mellifera (Figure 7A). By comparison, we have tested the polymorphism rate between A. mellifera subspecies. We find more polymorphism between A. m. melifera and the three other subspecies tested (A. m. caucasia, A. m. ligustica and A. m. scutellata) than the one obtained among the different colonies from A. m. mellifera. These RFLP data were analyzed in a genetic distance matrix (Figure 7B).
the genome compared to other classes of highly reiterated sequences found in insects such as the beetle *Tenebrio molitor*, the fruit fly *Drosophila nasutoides* or the parasitoids *Diadromus pulchellus* and *Eupelmus vuilleti* where they occur in a large portion, respectively 50%, 60%, 15% and 25% of their genome (Davis and Wyatt 1989; Cordero-Stone and Lee 1976; Bigot, Hamelin and Periquet 1990). Nevertheless, this is in agreement with previous studies (Jordan and Brosemer 1974; Crain, Davidson and Britten 1976) showing that less than 10% of the honeybee genome is composed of moderately and highly repeated DNA sequences.

The presence of bands outside of the AluI and DraI ladder patterns could be explained by two hypothesis. (1) They belong to classes of highly reiterated sequences closely related to the 176-bp AluI monomer coming from a common ancestral sequence. (2) They come from recent modifications (insertion or partial duplication inside the AluI multimers) followed by amplification of some of the repeat units. We can notice that evaluation of copy number of this class of DNA is probably overestimated due to cross-hybridization with these possible other classes of highly reiterated sequences.

A comparative analysis of our consensus sequence has provided evidence for an internal subrepeat varying from 11 to 17 bp in length. The prototype sequence of this subrepeat, deduced from the consensus sequence, is 5′AAAGAT(CG)(AT)GGCGAGTTT3′, which itself could come from an ancestral palindromic decamer 5′AAAGAGTTT3′. In our all analyzed monomers this motif is, with variations, present seven times, some of the subrepeats being in tandem array. Such an arrangement within the basic units strongly argues for the evolution of the *A. mellifera* AluI tandemly repeats from a common ancestral sequence by amplification followed by mechanism of sequence divergence, as it has already been proposed for other classes of highly reiterated sequences (Brutlag 1980; John and Miklos 1979; Singer 1982).

The *A. mellifera* AluI tandemly repeats class is unusual by its homogeneity. All the monomers are very similar to each other, deviating from the consensus on an average of only 1.4%. Our data on *A. mellifera* may be compared with some observations on other highly tandemly repeated DNAs. A high homogeneity has been shown for other classes of highly reiterated DNA like in *D. melanogaster* (3.6%) and *T. molitor* (2%) where a great intermonomer DNA sequence conservation exists (Hsieh and Brutlag 1979; Davis and Wyatt 1989). In rat (Sealy et al. 1981), Ascaris (Müller et al. 1982), calf (Pages and Roizes 1984) and in brine shrimp (Crucés et al. 1986), the intermonomeric divergence in highly reiterated DNA is higher and they can be divided into variant classes. By contrast, there was no clustering of variants within the bovine repeats (Pech, Streeck and Zachau 1979); however, the overall variation level was also high (12%).

Analysis of the phylogenetic tree of Apis repeats indicates that the different sequenced variants showed
This unusual homogeneity found in *A. mellifera* could be explained by two hypotheses. First, some mechanism, like gene conversion, may be acting upon the genome of *A. mellifera* to maintain sequence homogeneity, as has been proposed for *T. molitor* (Davis and Wyatt 1989). Second, the cloned *AluI* tandemly repeats belong to a family that may have recently appeared in the genome of *A. mellifera*. During the "life" of the repeat, nucleotide changes and rearrangements do not seem to occur randomly. In some cases C to T and G to A transitions are more frequent than the reverse transitions and consequently the tandemly repeat DNA sequences tend to become enriched in AT during evolution as is shown for *T. molitor* (Ugarkovic, Plohl and Gamulin 1989). Moreover, some small DNA insertions without evident advantageous features may appear in one unit of repeat and extend to become conserved in all the repeats of the family. This was observed for tandemly repeat family in bovine (Rozes and Pages 1982) and in the parasitic wasp Diadromus collaris (Bigot, Hamelin and Periquet 1990; Roujas-Rousse, Bigot and Periquet 1993).

In the case of the *AluI* highly reiterated sequences in *A. mellifera*, the percentage of A + T content (58%) is not far off from the average for honeybee (63%) (Jordan and Brosem 1974), and no particular AT tract can be detected. Furthermore, no insertion conserved in all the repeat units has been observed. These two facts added to the observation that it is highly homogeneous and constitutes only 2% of the genome suggest that the *AluI* repeats appeared recently in the *A. mellifera* genome.

However, the presence of the *AluI* tandem repeat family in *A. cerana* and its absence in *A. dorsata* strongly suggests that this family appeared after the divergence between *A. dorsata* and the common ancestor of *A. cerana* and *A. mellifera*, but before the *A. cerana*- *A. mellifera* cladogenese. According to mtDNA data (Garnery et al. 1991), this occurred around 6–9 myr ago. If now, we consider that as for Drosophila, the rate of nuclear and mitochondrial DNA divergences are equal (Powell et al. 1986; Caccone, Amato and Powell 1988), i.e., 2% per myr, the 2.5% divergence among the 10 clones, would lead to a much more recent origin for the *AluI* tandem repeat family, i.e., 1.25 myr. This situation could be explained by a recent amplification in *A. mellifera* after the split of *A. mellifera* and *A. cerana* lineages. In this case, *AluI* repeats should be less abundant in *A. cerana* than in *A. mellifera*. We found that this sequence has the same abundance in the two closely related species (data not shown). Therefore this result strongly argues for a mechanism acting upon the genome of *A. mellifera* to maintain homogeneity in this sequence rather than two amplification events taking place in genomes of
A. mellifera and A. cerana after speciation. This scenario is based on a similar rate of nucleotide substitution in mitochondrial and nuclear DNA. However, in the absence of any calibration concerning the nuclear DNA evolutionary rate in the Apis taxon, the second hypothesis previously proposed cannot be totally rejected.

In the consensus sequence, the presence of dyad intranastand structures induced by the palindrome could indicate a potential role for this highly reiterated sequence in the heterochromatin condensation process. The dyad structures distributed along the sequence of insect highly reiterated sequence DNA could play the same role as methylation signals do in the H1 histone in vertebrates (Davie and Delcuve 1988; Pages and Roizes 1988; Bigot, Hamelin and Periquet 1990). However, the fact that a highly reiterated sequence has dyad structures does not mean that it has a function and these structures could be the result of duplication in tandem repeats.

In Northern blot analysis, we failed to detect any transcript corresponding to our AluI satellite. Furthermore, the examination of the open reading frames indicates that translation of possible transcripts would stop codons. This situation corresponds to most of the transcription studies conducted on satellite DNAs. However, in some cases, highly reiterated classes of DNA were found to be transcribed in the newt lampbrush chromosomes (Varley, MacGregor and Erba 1980) or in rat hepatoma tissue culture cells (Sealy et al. 1980). This was generally interpreted as an accidental transcription from a promoter of adjacent genes (Varley, MacGregor and Erba 1980; Diaz et al. 1981).

Turning now to the prospect of polymorphism for analysis of within species variation, we have shown that AluI repetitive sequence is polymorphic within A. m. mellifera subspecies. This is shown for individuals from three different colonies, while the difference between individuals of the same colony is very weak. However we can notice that the polymorphism observed could be due to the AluI repetitive DNA itself or to sequences that cross-hybridize with it, some of the bands being outside of the typical ladder obtained with other restriction enzymes (data not shown).

So, the AluI highly reiterated sequences provide a new probe for fingerprinting in A. m. mellifera. Fingerprints have already been obtained in honeybees with minisatellites (Blanchetot 1991) and microsatellites (Moritz, Meusel and Haberl 1991) probes. Our data show that similar results can be obtained with large tandem repeat sequences. However, in small animals such as insects, fingerprinting through Southern blotting of genomic DNA necessitate generally most of the DNA that can be extracted from one individual. Polymerase chain reaction (PCR)-based markers are usually much more convenient, not only because they need minute amounts of DNA but also because the genetic interpretation of variation is generally much easier. Among these markers, microsatellites, which are currently under study (Estoup et al. 1993), appear to be the most promising. Nonetheless, because of their different locations on chromosomes and their putative variation in the mode of evolution, complementary information can be obtained from the various classes of tandem repeat sequences.

Although our data on polymorphism within A. mellifera are preliminary, the neighbor-joining tree based on RFLP distances (Figure 7B) is fully consistent with those obtained with mtDNA data (Cornuet and Garney 1991; Garney, Cornuet and Solignac 1992; Smith 1991a,b). A more thorough study might bring a valuable support to current hypothesis about the evolutionary history of honeybees.

We are grateful to Y. Leconte and B. Vaissière (INRA-Avignon) for providing honeybee lineages, C. Quiles for her technical assistance and F. Vanlerberge-Masutti for her critical discussions and helpful comments on the manuscript.

LITERATURE CITED


Honeybee Alu I DNA Sequence

Crozier, R. H., and Y. C. Crozier, 1993 The mitochondrial genome of the honeybee Apis mellifera: complete sequence and genome organization. Genetics 133: 97-117.


Communicating editor: A. G. Clark