A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis

ABSTRACT

Analysis of the currently available bacterial genome sequences classifies Bacillus anthracis and Yersinia pestis as having an average (approximately 30 per Mb) density of tandem repeat arrays longer than 100 bp when compared to the other bacterial genomes analysed to date. In both cases, testing a fraction of these sequences for polymorphism was sufficient to quickly develop a set of more than fifteen informative markers, some of which show a very high degree of polymorphism. In one instance, the polymorphism information content index reaches 0.82 with allele length covering a wide size range (600-1950 bp), and nine alleles resolved in the small number of independent Bacillus anthracis strains typed here.

INTRODUCTION

Background The polymorphism associated with tandem repeats has been instrumental in mammalian genetics for the construction of genetic maps and still is the basis of DNA fingerprinting in forensic applications. Tandem repeats are usually classified among satellites (spanning megabases of DNA, associated with heterochromatin), minisatellites (repeat units in the range 6-100 bp, spanning hundreds of base-pairs) and microsatellites (repeat units in the range 1-5 bp, spanning a few tens of nucleotides). More recently, a number of studies have supported the notion that tandem repeats reminiscent of mini and microsatellites are likely to be a highly significant source of very informative markers for the identification of pathogenic bacteria even when these pathogens are recently emerged, highly monomorphic species. This probably reflects the important contribution of tandem repeats to the adaptation of the pathogen to its host. Tandem repeats appear to contribute to phenotypic variation in bacteria in at least two ways. Tandem repeats located within the regulatory region of a gene can constitute an on/off switch of gene expression at the transcriptional level. Similarly, tandem repeats within coding regions with repeat units length not a multiple of three can induce a reversible premature end of translation when a mutation changes the number of repeats (reviewed in). In other instances, the repeated unit length is a multiple of three, and the tandem repeat contributes to a coding region. In such cases, variations in the number of copies modify the gene product itself. Mutation mechanisms of micro and minisatellites have been studied in some detail in eukaryotes, essentially human and yeast (reviewed in). In brief, the data obtained so far suggest that microsatellites mutate by replication slippage processes; mutation rates depend upon the efficiency of mismatch repair mechanisms and an internal heterogeneity within the array strongly stabilizes the tandem repeat. In contrast, minisatellites mutate predominantly as the result of the repair of a double strand break initiated within, or very close to, the tandem repeat. In eukaryotes at least, these events can be of replicative origin, or can be genetically controlled, and specifically induced, during meiosis, at double strand breaks hot-spots. Minisatellite mutation rate in eukaryotes appears to be insensitive to mismatch repair efficiency, and internal heterogeneity is compatible with a high mutation rate. In bacteria, loci containing a tandem repeat from the microsatellite class (repeat unit sizes of 1-8 bp) have been called simple sequence contingency loci. Altered number of repeats allows for reversible on and off states of expression for the corresponding gene. The mutation rate of a tetranucleotide (microsatellite) tract in Haemophilus influenzae is higher than 10-4 and contributes to the adaptation of the pathogen to its

hosts as the infection progresses. In such an extreme situation, the microsatellite is of limited value for strain identification, epidemiological and phylogenetic studies. The tandem repeat array is composed of perfect copies of the elementary unit, and different alleles are observed in a single culture. In contrast, the phylogenetic identity of minisatellite alleles of identical size can usually be further checked by DNA sequencing, since the repeated units are often not perfect. The pattern of variants along the array provides an additional level of allele identification and phylogenetic information. In addition, tandem repeats with longer repeat unit length can be relatively easily typed in the size range of a few hundred base-pairs using ordinary horizontal gel electrophoresis. In this report, we will first describe the use of a tandem repeats database for bacterial genomes () and briefly compare the general characteristics of tandem repeats in a number of bacterial genomes for which the sequence has been determined and made publicly available. We will then show how this tool can easily be applied to the rapid characterization of new highly polymorphic markers in two pathogens, Y. pestis and B. anthracis. Both Y. pestis (causative agent of plague) and B. anthracis (causative agent of anthrax) are recently emerged clones of respectively Y. pseudotuberculosis and B. cereus. In the case of Y. pestis, a high resolution typing tool based on RFLP (Restriction Fragment Length Polymorphism) analysis of IS100 locations has already been developed. However this technology is more demanding than PCR typing, which justifies the development of such an assay. In the case of B. anthracis, polymorphisms were initially identified essentially using AFLP (Amplified Fragment Length Polymorphism) typing. Subsequent analyses demonstrated that the most informative fragments in AFLP patterns resulted from tandem repeat array length variations (five minisatellite loci were characterized in this way).

CONCLUSION

Conclusions We limited here our investigation of tandem repeats to minisatellites, i.e. repeat units longer than 9 base-pairs, so as to avoid simple sequence contingency loci of limited epidemiological value, and to facilitate the typing of alleles with agarose gel electrophoresis. However, simple sequence contingency loci are also represented in the database and are of great interest for molecular pathogenicity studies. The use of the tandem repeats database was demonstrated here on two of the most genetically homogeneous human pathogenes, Y. pestis and B. anthracis. There is consequently a possibility that a common database format for identification and epidemiological analyses of pathogens amenable to minisatellite typing be developed. As more data becomes available on polymorphism associated with tandem repeats, it will be added to the database presented here in order to avoid duplication of work and nomenclature. Bacterial species differ very significantly in the density of tandem repeats within their genome, and also in their use of tandem repeats. Some species have a very strong excess of tandem repeats with repeat units length which are multiple of three, the most striking examples being M. tuberculosis and P. aeruginosa. Polymorphism in such tandem repeats is likely to modulate the protein structure rather than gene activity. In M. tuberculosis, all tandem repeats with total length (L) higher than 100 bp and 9 or 15 base-pairs long units are located with ORFs. An important proportion of these tandem repeats correspond to the so-called PE and PPE multigene families. In the two species studied here, tandem repeat polymorphism is strongly correlated with one or more of the sequenced allele characteristics, as illustrated in Figure 7. In Yersinia pestis a strong correlation is observed between number of alleles observed and homogeneity of the tandem array. In Bacillus anthracis, the strongest

correlations are with total array length and GC content. It appears that the correlations are not the same in the two species, so that at present at least, the polymorphism associated with a tandem repeat cannot be inferred from its primary sequence. In particular, and in contrast to what is known for microsatellites (1-5 bp repeat units), some of the minisatellites are highly polymorphic in spite of a poor internal homogeneity of the sequenced allele, as is also the case for minisatellites in the human genome. However, more systematic allele sequencing will be required to demonstrate that polymorphism is not associated with a subclass of alleles showing a higher internal homogeneity. Similarly, allele sequencing will be required to formally establish that the allele size variations observed are indeed (as is likely) the consequence of variations in the number of repeats. Five among the B. anthracis markers described here (Ceb-Bams1, 3, 7, 13 and 30) are highly polymorphic with PIC values (or Nei's index) above 0.7. In this respect, it is important to observe that the length of the allele observed for Ceb-Bams1 in the Ames strain is not of the size expected from the sequence data (Table 2). This may result either from a high mutation rate at Ceb-Bams1 or from a sequencing error. The expected allele size corresponds to allele 4 (Table 3), which is unlikely for the Ames strain because Ceb-Bams1 allele 4 is observed only in cluster B strains (Figure 6) and Ames is well apart of cluster B. A similar situation is observed for Ceb-Bams28, for which the expected product does not correspond to any existing allele in the collection of strains typed. In this case however, the locus is moderately polymorphic, with a PIC value of 0.26 and only three alleles observed (Table 2), so that a sequencing error is the most likely interpretation. This issue could be easily solved by typing with Ceb-Bams1 and Ceb-Bams28 the very strain which has been used for the sequencing project. It is interesting to observe that, although the magnitude of allele size difference has not been taken into account when building the distance matrix, the resulting phylogenetic tree proposed in Figure 6 tends to group together strains with alleles of similar size at these most variable loci. This is reminiscent of observations made in H. influenzae and suggest that mutation events are predominantly small size changes. Here again, more detailed studies involving full allele sequencing should now help understand the succession of events producing a population of alleles.