Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria

Nigel J. Robinson*, Pamela J. Robinson, Amit Gupta, Alan J. Bleasby¹,
Brian A. Whitton² and Andrew P. Morby⁺

Department of Biochemistry and Genetics, The Medical School, University of Newcastle, Newcastle NE2 4HH, UK, ¹Daresbury Laboratory, Daresbury, Warrington WA4 4AD, UK and ²Department of Biological Sciences, University of Durham, Durham DH1 3LE, UK

Received December 19, 1994; Revised and Accepted January 26, 1995

ABSTRACT

An octameric palindrome (5'-GCGATCGC-3') is abundant in cyanobacterial sequences within databases (GenBank/EMBL) and was designated HIP1 (highly iterated palindrome). The frequency of occurrence of all 256 octameric palindromes has now been determined in sub-databases revealing large and unique over-representation of HIP1 in cyanobacterial entries. DNA sequences from other bacteria were searched for any over-represented octameric palindromes analogous to HIP1. Only two sequences were identified, in the genomes of a thermophile and halophilic archaeobacteria, although these were less abundant than HIP1 in cyanobacteria and relate to codon usage. To test the proposed widespread distribution of HIP1 in DNA from the cyanobacterium *Synechococcus* PCC 6301, randomly selected genomic clones were partly sequenced. HIP1 constituted 2.5% of the novel sequences, equivalent to a site on average once every 320 nucleotides. An oligonucleotide including HIP1 was also tested in PCR. Multiple products were obtained using template DNA from cyanobacterial strains in which HIP1 is abundant in known sequences, and some strains generated characteristic HIP-PCR banding patterns. However, analysis of DNA from one strain (not previously represented in databases) by random sequencing, HIP-PCR and *PvuI* digestion, confirms that not all cyanobacterial genomes are rich in HIP1.

INTRODUCTION

Repetitive DNA elements are known to be dispersed throughout many genomes, the *Alu* family of repeats in mammals being one of the more extensively studied (1–4). Repetitive DNA sequences in bacterial genomes include: two classes of bacterial interspersed mosaic elements (BIME-1 and BIME-2) (5) of 40–500 nucleotides (nt) in length and found in the genomes of several members of the Enterobacteriaceae; a ~126 nt enterobacterial repetitive

intergenic consensus/intergenic repeat unit (ERIC/IRU) (6,7) found in many Gram-negative bacteria (8); three classes of short tandemly repeated repetitive (STRR) elements found in some cyanobacteria (9); BOX sequences of up to 280 nt found in *Streptococcus* spp. (10); Bru-RS1 and Bru-RS2 of 103 and 105 nt found in *Brucella* sp. (11); RMX elements with an 87 bp core found in *Myxococcus xanthus* (12) and SARK, a family of palindromic repeats in *Deinococcus radiodurans* which vary in length from 150 to 192 nt (13). Both BIME and BOX sequences consist of different short conserved motifs which in BIME includes a ~40 nt repetitive extragenic palindrome (REP) element of dyad symmetry (14), also referred to as PU (palindromic unit) (15–17), and in the BOX sequences includes boxA, boxB and boxC of 59, 45 and 50 nt, respectively (10). All of these small conserved sequences are dissimilar to known transposable elements and the bases of proliferation and conservation remain to be established. Examples of functions attributed to some of these sequences include roles in the termination, stability and regulation of translation of certain transcripts (18–21), the organisation of bacterial chromosome structure (15,16,21) and the promotion of genomic rearrangements (16,22). With the amassing of DNA sequence information, computer and statistical methods are now being applied to the identification of additional significantly recurring sequences within genomes (23).

HIP1 was first identified in association with a gene deletion in mutants of *Synechococcus* PCC 6301 (24). In cells selected to grow in Cd²⁺, part of the *smtB* gene, encoding a repressor of expression of cyanobacterial metallothionein, was lost and the palindromic sequence 5'-GCGATCGC-3' traversed the deletion termini (24). This sequence is present seven times within the 1.3 kb *smt* divergon and was found to be common in other sequences from cyanobacteria within GenBank (and EMBL) release 70.0 (24). Within *smt*, HIP1 sites occur both inside (four sites) and outside (three sites) protein coding regions, and (at least in *smt*) these sites are not clustered in adjacent repeats. These features suggest that HIP1 is distinct from any known interspersed repetitive DNA sequence elements. The frequency of iteration of HIP1 in *Synechococcus* strains was estimated to be once every ~664 nt within sequences in GenBank (and EMBL) release 70.0

* To whom correspondence should be addressed

⁺Present address: School of Biological Sciences, University of Birmingham, Birmingham B15 2TT, UK

(24). This is clearly higher than the predicted occurrence of this 8 bp sequence [once every 45 670 nt for organisms with 55% G+C; G+C content for *Synechococcus* PCC 6301 is estimated to be 55.1% (25)], but is this a substantial departure from the frequencies of occurrence of all other octameric palindromes within known cyanobacterial DNA? Furthermore, are there equally abundant octameric palindromes found in the DNA of many other bacteria?

In this report we establish the singular nature of HIP1 by determining the frequency of occurrence of all 256 octameric palindromes within sequences in the GenBank (and EMBL) databases. Of course, it cannot be assumed that known gene sequences represent a random sample of a chromosome as a whole, for example many characterised cyanobacterial genes are involved in either photosynthesis or nitrogen fixation and database searches may thus be distorted to reflect unique properties of these genes. Therefore, attempts have been made to estimate the frequency of occurrence of HIP1 experimentally both in organisms represented in the databases and in cyanobacterial isolates from which DNA sequences have not previously been reported. Finally, repetitive DNA sequences have previously been exploited in the design of PCR primers (4,8) which have been used to generate species- and strain-specific DNA fingerprints for some bacteria (8). The use of HIP-PCR to analyse cyanobacterial DNA is also described.

MATERIALS AND METHODS

Materials, oligonucleotides and DNA sequencing

DNA restriction and modifying enzymes were supplied by Boehringer Mannheim Ltd (Lewes, UK) and Northumbria Biologicals Ltd (Cramlington, UK). An oligonucleotide used for *in vitro* DNA amplification in HIP-PCR was synthesised using a 381A DNA synthesiser (Applied Biosystems Ltd, Warrington, UK) according to the manufacturer's protocol. Plasmid sequencing was performed by the dideoxy-method (26) and reaction products analysed using a 370A DNA sequencer (Applied Biosystems Ltd, Warrington, UK).

Cyanobacterial strains and DNA isolates

Calothrix D253 (isolated from a mangrove in Cuba obtained originally from J. Komarek, Trebon, and from which no DNA sequences had previously been reported); *Calothrix* PCC 7601 (*Fremyella diplosiphon*); *Synechococcus* PCC 6301 and *Synechocystis* PCC 6803; *Synechococcus* PCC 7942 (strain R2-PIM8) were cultured in AD containing Fe (0.4 mg/l) and PO₄ (1 mg/l) (27); modified BG11 (28); modified AC (29,30); and Allen's (31) with 0.012 g/l citric acid, the omission of Na₂SiO₃·9H₂O and supplemented with 5 µg/ml streptomycin and 30 µg/ml DL-methionine media, respectively. Cyanobacterial cultures were grown under constant light (100 µmol photon m⁻² s⁻¹ photosynthetically active radiation) at 32°C. Genomic DNA from these cyanobacterial strains was purified (32) using CsCl gradients. Genomic DNA from other cyanobacterial strains was kindly provided by D. G. Adams (Leeds University, UK; *Synechococcus* PCC 7942, *Synechocystis* PCC 6803, *Nostoc* LBG1, 2A5, 2A6, 4A3, 4A4 and 1B6), N. H. Mann (Warwick University, UK; *Synechococcus* PCC 7942, PCC 7002, WH 8012, WH 8018, WH 8103 and WH 7803, *Synechocystis* PCC 6803, *Nostoc* MAC PCC 8009), D. Valout (Roscoff Biological

Station, France; marine strain *Synechococcus* Ros 04), A. R. Grossman (Carnegie Institution of Washington, USA; *Synechococcus* PCC 7942, *Synechocystis* PCC 6803 and PCC 6701), H. B. Pakrasi (Washington University in St Louis, USA; *Synechocystis* PCC 6803), P. S. Bisen (Barkatullah University, India; *Nostoc* MAC and MAC R1), K. Forsland and P. Lindblad (Uppsala University, Sweden; *Nostoc* PCC 8005, PCC 8002 and PCC 7120).

Genomic libraries and PCR

Libraries of *Sau*3AI fragments of partially digested genomic DNA from *Calothrix* PCC 7601, and *Eco*RI fragments of genomic DNA from *Calothrix* D253, both in λ ZAPII (Stratagene Ltd, Cambridge, UK), were provided by N. Federspeil (Microcide Pharmaceuticals, USA) and I. M. Evans (Durham University, UK), respectively. Phagemid pBluescript was recovered from randomly selected clones and used as sequencing template. A genomic library of *Pst*I-restricted genomic DNA from *Synechococcus* PCC 6301 was generated in pGEM-3Zf+ (Promega Ltd, Southampton, UK) and plasmid DNA purified from randomly selected transformants was used as sequencing template.

In vitro DNA amplification was performed using an Omnigene Temperature Cycler (Hybaid Ltd, Teddington, UK) with 36 cycles of denaturation (94°C, 1 min), annealing (55°C, 1 min) and extension (72°C, 1 min 30 s) followed by 1 cycle of denaturation (94°C, 1 min), annealing (55°C, 1 min) and extension (72°C, 6 min) and standard reaction conditions as used previously (24). An aliquot (70 µl) of each 100 µl reaction was analysed by electrophoresis on 1% agarose gels stained with ethidium bromide.

Computer analyses

Computer sequence analyses were performed on either the Daresbury facility DLVH using the UWGCG (24) and PIR programs or using UWGCG on the SEQNET facility at the Daresbury Laboratory (DRAL). Data was collected from the bacterial entries within GenBank (and EMBL) using either release 74.0 or release 85.0, as indicated, dependent upon when the search was performed. Sub-databases containing all sequence entries for either a particular organism, or group of organisms, were generated by identifying, recalling and compiling entries defined by the name recorded in the databank under the organism designation using STRINGSEARCH within the UWGCG package of programs. The sub-databases were manually edited to avoid duplications and ambiguities. Subsequently, a nucleotide sequence database for each generic entry was compiled using the FETCH and TOPIR programs. Nucleotide sequence databases were similarly compiled for individual species within a genera. The derivative databases were searched for the frequency of occurrence of all possible hexameric and octameric palindromes using the PAL1 program written for this work by AJB and now available within the SEQNET service. This programme may be accessed by typing 'pal1' (lower case characters) within UWGCG and the number of copies of each possible palindrome of a defined size (half the number of nucleotides to be specified) in a sequence file determined using the following command line: pal1_[topir file name]_[half the no. of nucl.]>_[output file name].

Table 1. The estimated average distance (in nucleotides) between HIP1 sites and between *PvuI* sites within DNA from individual cyanobacterial species for which there is at least 10 kb of known sequence (amount shown) in GenBank (and EMBL) release 85.0

Organism	HIP1	<i>PvuI</i>	Known sequence
<i>Synechococcus</i> PCC 6301	562	300	47742
<i>Synechococcus</i> PCC 7942	479	260	15 3262
<i>Synechococcus</i> PCC 7002	785	692	47072
<i>Synechococcus</i> PCC 6716	608	584	15188
<i>Synechococcus</i> WH 8020	∞	2127	14891
<i>Synechococcus vulcanus</i>	1737	1274	19107
<i>Synechococcus elongatus</i>	843	674	20223
<i>Calothrix</i> PCC 7601	1398	908	34372
<i>Anabaena variabilis</i> PCC 7120	1609	1237	16 5695
<i>Anabaena variabilis</i> PCC 7937	1929	1286	27000
<i>Pseudanabaena</i> PCC 7409	805	499	10470
<i>Synechocystis</i> PCC 6803	1731	962	19 0403
<i>Plectonema boryanum</i> M101	11 506	397	11506

Calothrix PCC 7601 sequences include entries under the designation *Fremyella diplosiphon*, and entries under the designation *Anabaena variabilis* ATCC 29413 are included under *Anabaena variabilis* PCC 7937.

RESULTS

Comparison of the abundance of HIP1 with other octameric palindromes in known sequences from cyanobacteria and *E.coli*

To compare the abundance of HIP1 with that of other octameric palindromes, the frequency of occurrence of all 256 octameric palindromes was estimated in sequences within GenBank (and EMBL) release 74.0. HIP1 is considerably more abundant than any other octameric palindrome within a sub-database containing cyanobacterial sequences (Fig. 1A). For comparison, an equivalent analysis was performed on known sequences from *E.coli* where no individual octameric palindrome showed equivalent over-representation (Fig. 1B). Table 1 lists the estimated average distance between HIP1, and between *PvuI*, sites in cyanobacteria for which there is >10 kb of known sequence within GenBank (and EMBL) release 85.0. HIP1 is over-represented in most of the accumulated cyanobacterial sequences within the database. However, no sites were detected within known sequence (14 891 bp) from *Synechococcus* WH 8020 suggesting that DNA from some cyanobacteria may not contain abundant HIP1 sites.

Identification of any abundant octameric palindromes in other microbial groups

To identify any sequences similar to HIP1 which are abundant in known DNA from any other (than cyanobacteria) bacterial group, individual searches of GenBank (and EMBL) release 74.0 were performed for each of the 256 octameric palindromes. The output was analysed to identify multiple occurrence of a given palindrome within several database entries for a single species. Only two octameric palindromes, 5'-CGTCGACG-3' (designated HIP2) and 5'-CCTCGAGG-3' (designated HIP3), were identified in halobacteria and *Thermus* sp., respectively. Figure 2a and c shows the frequency of HIP2 and HIP3 in database entries for all microbial genera, or groups, for which there was at least 10 kb of

known sequence (with the exception of *Halococcus* sp. where only 9.2 kb was known) and one or more site.

Two sub-databases were created containing database entries for either halobacteria, or for *Thermus* sp., and these were searched for the abundance of different octameric palindromes (Fig. 2b and d). HIP2 is the most abundant of a group of three related sequences 5'-CG(A/C/T)CG(T/G/A)CG-3' (Fig. 2b).

Examination of HIP1 distribution and abundance in cyanobacterial DNA via random sequence analysis

To investigate whether or not the apparent over-representation of HIP1 in database entries is representative of a cyanobacterial genome, plasmids were purified from a library prepared from *PstI*-digested *Synechococcus* PCC 6301 genomic DNA and sequenced using either M13 forward (21 plasmids) or M13 reverse (19 plasmids) primers to obtain 14 423 bp of novel sequence. The frequency of occurrence of different octameric palindromes within these sequences confirms the singular over-representation of HIP1 previously noted in database entries for this organism (Fig. 3). A HIP1 site occurred on average once every 320 nt (*PvuI* sites were detected once every 162 nt). There were only two instances (four HIP1 sites) where sites were located within 20 nt of each other, and a further three (six HIP1 sites) within 50 nt of each other. The majority (83%) of degenerate (7 nt matching) HIP1 sites were mismatched at 'either' terminal nucleotide rather than the central hexanucleotide and this is partly reflected in the prevalence of *PvuI* sites noticeably exceeding HIP1. Sequence (8272 bp) analysis of inserts within plasmids recovered from a library prepared from *Sau3AI*-digested DNA from *Calothrix* PCC 7601 identified the sequence 5'-GATCGC-3' (*Sau3AI* site underlined) at the start of 62% of the inserts, substantially greater than the prediction that only 6.3% of the *Sau3AI* sites would be followed by the dinucleotide 'GC' and consistent with over-representation of HIP1 in *Calothrix* PCC 7601.

Figure 1. Frequency of occurrence of the 256 possible octameric palindromes in sub-databases of GenBank (and EMBL) release 74.0, reveals massive and singular over-representation of HIP1 in cyanobacteria. The first and second nucleotide pairs are listed in the order AA, CA, GA, TA, AC, CC, GC, TC, AG, CG, GG, TG, AT, CT, GT and TT from front to back and from left to right, respectively. (A) Analysis of a sub-database of cyanobacterial sequences. (B) Analysis of a sub-database of sequences from *E.coli*.

Plasmids were recovered from a library prepared with *EcoRI*-digested genomic DNA from *Calothrix* D253 and 19 301 bp of novel sequence generated. There were neither HIP1 sites nor *PvuI* sites within these sequences.

Estimation of HIP1 abundance using *PvuI* digestion and HIP-PCR

The recognition sequence for *PvuI* (underlined) is contained within HIP1 (5'-GCGATCGC-3'). It is predicted (Table 1) that many cyanobacteria contain HIP1 (and *PvuI*) sites more often than once every 4096 nt (4^6 , corresponding to an average frequency of occurrence for a 6 bp palindrome). To test this prediction DNA from *E.coli*, *Calothrix* D253, *Synechococcus* PCC 7942 and PCC 6301 was digested with *PvuI* and relatively (compared to *E.coli*) small fragments were generated for *Synechococcus* PCC 6301 and PCC 7942 (Fig. 4A). For *Synechococcus* PCC 6301 this is consistent with the observation that *PvuI* sites occurred once every 162 nt, twice the frequency of the HIP1 site, within novel sequences. DNA from *Calothrix* D253 was refractory to *PvuI*, consistent with an observed absence of either HIP1 or *PvuI* sites in the sequences obtained from this organism and hence supporting a conclusion that HIP1 and the

PvuI recognition site are uncommon in *Calothrix* D253. It is noted that this DNA sample was susceptible to digestion with other restriction endonucleases, *Sau3AI* and *EcoRI* (the latter being used in the production of the *Calothrix* D253 library).

Analysis of PCR products obtained using 1.5 min extension times, a primer (5'-GCCIIIGCGATCGC-3', I = inosine) containing the HIP1 sequence (underlined) and template DNA (100 ng) from either *Synechococcus* PCC 6301 or PCC 7942, produced a smear with the majority of the products <1.5 kb in length (Fig. 4B). In addition, bands, although faint, are apparent in the products obtained with *Synechococcus* PCC 7942 and PCC 6301 (Fig. 4B). By contrast, no products were detected with template DNA from either *E.coli* or *Calothrix* D253 suggesting an absence of adjacent HIP1 sites which can be 'traversed' by *Taq* polymerase (under these reaction conditions) in these organisms. By analogy, *Alu*-PCR has been performed with primers specific to primate *Alu* repeats and template DNA from somatic cell hybrids containing various numbers of human chromosomes within rodent cell backgrounds (4). It was noted that there was no strict correlation between the amount of human DNA and the number of bands amplified, but a background smear was more pronounced for hybrids with multiple human chromosomes, suggesting that sufficient fragments were amplified from cells containing abundant human *Alu* sequences to obscure individual products (4).

HIP-PCR can generate strain-specific DNA fingerprints for some cyanobacteria

To investigate the possibility that, by analogy to REP-PCR and ERIC-PCR (8), HIP-PCR may be used to generate strain-specific fingerprints for some cyanobacteria, reactions were repeated with template DNA from different cyanobacterial strains. To test the reproducibility of any observed banding patterns, replicate DNA samples isolated from independent cultures (by different research groups) were analysed for some strains.

A smear was obtained using DNA from several strains (Fig. 5), including *Synechococcus* PCC 7002 and *Synechocystis* PCC 6803 for which searches of GenBank (and EMBL) release 85.0 confirm that HIP1 occurs frequently in known DNA sequences (Table 1). In addition there are some prominent bands, and for *Synechococcus* PCC 7942 (Fig. 5A), *Synechocystis* PCC 6803 (Fig. 5B) and *Nostoc* MAC (Fig. 5C) the banding patterns are reproducible for each strain. For some strains, such as the novel *Nostoc* isolates designated LBG1, 2A5, 2A6 and 4A3 from thalli of the bryophyte *Anthoceros* (Fig. 5D), the smear is less intense suggesting that HIP1 may be somewhat less abundant in these strains, but characteristic banding patterns are apparent.

DISCUSSION

In comparison to all other octameric palindromes, HIP1 is substantially over-represented in collated database entries for cyanobacteria (Fig. 1). A search of DNA sequences from other bacteria for any abundant octameric palindromes analogous to HIP1 identified only two sequences in some halobacteria and *Thermus* sp. (Fig. 2). These sequences were not estimated to be as abundant (in database entries from these bacteria) as HIP1 (within database entries from cyanobacteria). Future studies could involve searching for palindromes differing in length to HIP1 which may be over-represented within certain bacterial genomes. Different mechanisms may, of course, be responsible for the proliferation and conservation of HIP1, HIP2 and HIP3 in

Figure 2. Estimated frequency of occurrence of HIP2 (5'-CGTCGACG-3') (a) and HIP3 (5'-CCTCGAGG-3') (c) in different bacterial genera indicates over-representation in certain halobacteria and *Thermus* sp., respectively. The frequency of occurrence of each palindrome is shown per nucleotide of sequence information for different genera (or groups) for which there was >10 kb of known sequence in GenBank (and EMBL) release 74.0 (with the exception of *Halococcus* sp. for which 9.2 kb was known) and at least one site. The frequency of occurrence of all 256 octameric palindromes in halobacteria (b) or *Thermus* sp. (d) in respective sub-databases, showing the relative abundance of HIP2 and HIP3. The first and second nucleotide pairs are listed in the order AA, CA, GA, TA, AC, CC, GC, TC, AG, CG, GG, TG, AT, CT, GT and TT from front to back and from left to right, respectively.

the respective bacterial genomes. Indeed, further examination of the distribution of these sites reveals that most HIP2 and HIP3 sites occur within protein coding regions and are found in a single reading frame (>80% of sites) encoding acidic amino acids (VD and LE, respectively) which are prevalent in proteins from these organisms. By contrast, examination of the context of HIP1 sites in *Synechococcus* sp. entries within GenBank (and EMBL) release 74.0 revealed sites on average once every 653 and 340 nt within coding and non-coding sequences, respectively. Furthermore, those HIP1 sequences which occur within protein coding regions are found in all three possible reading frames (24).

There was concern about the validity of extrapolating from the frequencies of occurrence of HIP1 within database entries to frequencies within entire genomes since database sequences may

be atypical. However, three experimental observations now confirm widespread over-representation of HIP1 in DNA from some cyanobacteria. First, genomic DNA from *E. coli* is more refractory to *PvuII* than DNA from *Synechococcus* PCC 7942 and PCC 6301 (Fig. 4A). Secondly, template DNA from each of these cyanobacteria supported the synthesis of more PCR products, using HIP1 containing primers, than DNA from *E. coli* (Fig. 4B). Thirdly, a HIP1 site occurs on average once every 320 nt within novel sequence generated from a *Synechococcus* PCC 6301 library (Fig. 3). Singular over-representation of HIP1 compared to all other octameric palindromes is also confirmed within these novel sequences (Fig. 3). Differences in HIP1 frequencies estimated from database entries for *Synechococcus* PCC 6301 (Table 1) and from random sequencing could reflect small sample

Figure 3. Frequency of occurrence of 256 octameric palindromes in new sequences (14 423 bp) from clones selected at random from a *Synechococcus* PCC 6301 genomic library. The first and second nucleotide pairs are listed in the order AA, CA, GA, TA, AC, CC, GC, TC, AG, CG, GG, TG, AT, CT, GT and TT from front to back and from left to right, respectively.

Figure 4. Digestion of genomic DNA from four organisms with *PvuI* and PCR using primers containing HIP1. Genomic DNA was either incubated with *PvuI* (A) or used as template in PCR reactions (B). Products were resolved on 1% agarose gels.

sizes and an uneven distribution of sites, and/or variation in HIP1 abundance in different cultures of this organism.

Initial analyses of sequences from those organisms represented in GenBank (and EMBL) release 74.0 suggested that HIP1 may be abundant in all cyanobacterial genomes, but not in other

Figure 5. HIP-PCR using different cyanobacterial DNA isolates. For some strains replicate DNA samples were analysed, each replicate having been obtained from a different laboratory (refer to Materials and Methods for sources), and products were resolved on 1% agarose gels. (A) *Synechococcus* PCC 7942 from four separate sources, *Synechococcus* PCC 7002 and *Synechococcus* Ros 04. (B) *Synechocystis* PCC 6803 from five separate sources and *Synechocystis* PCC 6701. (C) *Nostoc* MAC from three separate cultures and two sources, *Nostoc* PCC 8005, PCC 8002 and PCC 7120. (D) Novel 'local' isolates of *Nostoc*.

microorganisms (24). However, Table 1 shows that in the marine cyanobacterium, *Synechococcus* WH 8020, no HIP1 sites occur within sequences in GenBank (and EMBL) release 85.0. No products were detected following HIP-PCR using DNA from marine *Synechococcus* strains WH 8012, WH 8018 and WH 8103, and only a single high M_r product was detected with *Synechococcus* WH 7803 DNA (data not shown). It will be of interest to establish whether or not HIP1 over-representation is uncommon (or absent) in marine cyanobacteria. DNA from *Calothrix* D253, a strain from which no genes had previously

been sequenced, was refractory to *Pvu*I (Fig. 4A), did not support the production of abundant HIP-PCR products (Fig. 4B), and no HIP1 sites were detected in 19 301 bp of novel sequence from this cyanobacterium. The absence of detected HIP-PCR products using DNA from several other strains (Fig. 5) could also indicate that HIP1 is not abundant in these cyanobacteria. Since HIP1 is not abundant in all cyanobacterial genomes, an extensive analysis of the species distribution of HIP1 over-representation may assist in the identification of factors (e.g. parasitic or functional) responsible for the proliferation and conservation of this site within many of these genomes.

Similarities in the HIP-PCR banding patterns obtained using multiple isolates of DNA (isolated by several independent research groups) from the same strains suggest that this method can also be used to generate strain-specific fingerprints for some cyanobacteria (Fig. 5). HIP-PCR products obtained from two (2A5 and 2A6) of six novel *Nostoc* strains were similar (Fig. 5D). We have subsequently been informed that 2A5 and 2A6 were isolated from the same *Anthoceros* plant and that PCR using arbitrary primers has also suggested that these two strains are closely related (N. J. West, University of Leeds, personal communication). It is noted that for one of the three *Nostoc* MAC samples a high M_r product is more prominent while a low M_r product is less apparent (Fig. 5C). It is possible that subtle differences in the banding patterns for an organism might be indicative of gene rearrangements in the different lines maintained in several laboratories. *Nostoc* MAC is known to have undergone phenotypic changes in culture generating non-nitrogen fixing strains from which revertants (designated *Nostoc* MAC R) were obtained (33).

In addition to the singular nature of HIP1 over-representation compared to other octameric palindromes (Figs 1 and 3), the combination of several other features including its small size (8 bp, generally not clustered in adjacent repeats), perfect palindromic structure, very high frequency of iteration (estimated to occur on average once every 320 nt in *Synechococcus* PCC 6301), phylogenetic distribution (apparently restricted to some cyanobacteria) and localisation (including sites within protein coding regions) also distinguish HIP1 from other known prokaryotic repetitive DNA sequences. Furthermore, we have been unable to identify any other abundant octameric palindrome within another bacterial group or genus which is (fully) analogous to HIP1. The basis(es) of HIP1 propagation is(are) unknown, but the very high HIP1 occurrence seen in *Synechococcus* PCC 6301 and its common location within protein coding regions would be expected to represent a considerable burden if it did not confer some selective advantage. It is tempting to speculate that this site enhances genome plasticity, increasing the pool of genetic diversity, since it was first identified at the borders of a gene deletion (24). Excessive gene rearrangement would be disadvantageous and therefore the very high frequency of occurrence of HIP1 might suggest that a widespread role in gene rearrangement is unlikely. However, cyanobacteria are noted for containing multiple chromosomes with up to 30 copies per cell being reported (34). It is hypothesised that such polyploidy may accommodate widespread unfavourable HIP1-mediated gene-rearrangements within cyanobacterial cells.

ACKNOWLEDGEMENTS

This work was supported by research grant GR3/8660 from the UK Natural Environment Research Council. NJR acknowledges

support from the Royal Society and BAW from grant GST/02/574 from the UK Natural Environment Research Council. We wish to thank all the individuals who provided genomic DNA samples and thank Dave Scanlan, Nyree West and Jeff Elhai for most helpful advice. This work benefited from the use of the SEQNET facility.

REFERENCES

- Schmid, C.W. and Jelinek, W.R. (1982) *Science*, **216**, 1065–1070.
- Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Biro, P.A., Choudary, P.V., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L. and Schmid, C.W. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 1398–1402.
- Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1979) *J. Mol. Biol.*, **132**, 289–306.
- Nelson, D.L., Ledbetter, S.A., Corbo, L., Victoria, M.F., Ramírez-Solis, R., Webster, T.D., Ledbetter, D.H. and Caskey, C.T. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 6686–6690.
- Bachelier, S., Saurin, W., Perrin, D., Hofnung, M. and Gilson, E. (1994) *Mol. Microbiol.*, **12**, 61–70.
- Hulton, C.S.J., Higgins, C.F. and Sharp, P.M. (1991) *Mol. Microbiol.*, **5**, 825–834.
- Sharples, G.J. and Lloyd, R.G. (1990) *Nucleic Acids Res.*, **18**, 6503–6508.
- Versalovic, J., Koeuth, T. and Lupski, J.R. (1991) *Nucleic Acids Res.*, **19**, 6823–6831.
- Mazel, D., Houmard, J., Castets, A.M. and Tandeau de Marsac, N. (1990) *J. Bacteriol.*, **172**, 2755–2761.
- Martin, B., Humbert, O., Camara, M., Guenzi, E., Walker, J., Mitchell, T., Andrew, P., Prudhomme, M., Alloing, G., Hakenbeck, R., Morrison, D.A., Boulnois, G.J. and Claverys, J.P. (1992) *Nucleic Acids Res.*, **20**, 3479–3483.
- Halling, S. and Bricker, B.J. (1994) *Mol. Microbiol.*, **14**, 681–689.
- Fujitani, S., Komano, T. and Inouye, S. (1991) *J. Bacteriol.*, **173**, 2125–2127.
- Lennon, E., Gutman, P.D., Yao, H. and Minton, K.W. (1991) *J. Bacteriol.*, **174**, 2137–2140.
- Stern, M.J., Ferro-Luzzi Ames, G., Smith, N.H., Robinson, E.C. and Higgins, C.J. (1984) *Cell*, **37**, 1015–1026.
- Gilson, E., Clément, J.M., Perrin, D. and Hofnung, M. (1987) *Trends Genet.* **3**, 226–230.
- Gilson, E., Clément, J.M., Brutlag, D. and Hofnung, M. (1984) *EMBO J.*, **3**, 1417–1421.
- Higgins, C.F., Ferro-Luzzi Ames, G., Clément, J.M. and Hofnung, M. (1982) *Nature*, **298**, 760–762.
- Gilson, E., Rousset, J.P., Clément, J.M. and Hofnung, M. (1986) *Ann. Inst. Pasteur*, **137**, 259–270.
- Newbury, S.F., Smith, N.H., Robinson, E.C., Hiles, I.D. and Higgins, C.F. (1987) *Cell*, **48**, 297–310.
- Newbury, S.F., Smith, N.H. and Higgins, C.F. (1987) *Cell*, **51**, 1131–1143.
- Stern, M.J., Prossnitz, E. and Ferro-Luzzi Ames, G. (1988) *Mol. Microbiol.*, **2**, 141–152.
- Shyamala, V., Schneider, E. and Ferro-Luzzi Ames, G. (1990) *EMBO J.*, **9**, 939–946.
- Blaisdell, B.D., Rudd, K.E., Matin, A. and Karlin, S. (1993) *J. Mol. Biol.*, **229**, 833–848.
- Gupta, A., Morby, A.P., Turner, J.S., Whitton, B.A. and Robinson, N.J. (1993) *Mol. Microbiol.*, **7**, 189–195.
- Herdman, M., Janvier, M., Waterbury, J.B., Rippka, R. and Mandel, M. (1979) *J. Gen. Microbiol.*, **111**, 63–71.
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
- Allen, M.B. and Arnon, D.I. (1955) *Plant Physiol.*, **30**, 366–372.
- Stanier, R.Y., Kunisawa, R., Mandel, M. and Cohen-Bazire, G. (1971) *Bacteriol. Rev.*, **35**, 171–205.
- Shehata, F.H.A. and Whitton, B.A. (1981) *Br. Phycol. J.*, **17**, 5–12.
- Kratz, W.A. and Myers, J. (1955) *Am. J. Bot.*, **42**, 282–287.
- Allen, M.M. (1968) *J. Phycol.*, **4**, 1–4.
- Gupta, A. (1992) Ph.D. thesis, Durham University, UK.
- Meeks, J.C., Campbell, E.L. and Bisen, P.S. (1994) *Microbiology*, **140**, 3225–3232.
- Mann, N.H. and Carr, N.G. (1974) *J. Gen. Microbiol.*, **83**, 399–405.