BMC Bioinformatics

BioMed Central

Research article

Identification of signature and primers specific to genus *Pseudomonas* using mismatched patterns of 16S rDNA sequences

HJ Purohit*, DV Raje and A Kapley

Address: Environmental Modeling and Genomics Division, National Environmental Engineering Research Institute, Nehru Marg, Nagpur – 440020 (MS) India

Email: HJ Purohit* - hemantdrd@hotmail.com; DV Raje - dv_raje@hotmail.com; A Kapley - atyadrd@hotmail.com

* Corresponding author

Published: 22 May 2003

BMC Bioinformatics 2003, **4**:19

Received: 5 November 2002
Accepted: 22 May 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/19

Abstract

Background: *Pseudomonas*, a soil bacterium, has been observed as a dominant genus that survives in different habitats with wide hostile conditions. We had a basic assumption that the species level variation in 16S rDNA sequences of a bacterial genus is mainly due to substitutions rather than insertion or deletion of bases. Keeping this in view, the aim was to identify a region of 16S rDNA sequence and within that focus on substitution prone stretches indicating species level variation and to derive patterns from these stretches that are specific to the genus.

Results: Repeating elements that are highly conserved across different species of *Pseudomonas* were considered as guiding markers to locate a region within the 16S gene. Four repeating patterns showing more than 80% consistency across fifty different species of *Pseudomonas* were identified. The sub-sequences between the repeating patterns yielded a continuous region of 495 bases. The sub-sequences after alignment and using Shanon's entropy measure yielded a consensus pattern. A stretch of 24 base positions in this region, showing maximum variations across the sampled sequences was focused for possible genus specific patterns. Nine patterns in this stretch showed nearly 70% specificity to the target genus. These patterns were further used to obtain a signature that is highly specific to *Pseudomonas*. The signature region was used to design PCR primers, which yielded a PCR product of 150 bp whose specificity was validated through a sample experiment.

Conclusions: The developed approach was successfully applied to genus *Pseudomonas*. It could be tried in other bacterial genera to obtain respective signature patterns and thereby PCR primers, for their rapid tracking in the environmental samples.

Background

Pseudomonas, a soil bacterium, because of its diverse catabolic potential is often associated with various microbial communities surviving in different environmental conditions. Hence, it is often considered as an indicator organism to study the development and dynamics of microbial community. We are working on treatment of wastewater having toxic chemicals with the emphasis on identifying bacterial isolates [1–6]. We have sequenced 16S rRNA gene for two strains; *Pseudomonas* SF1 (utilizes 4-nitrophenol) and PH1 (utilizes 3-aminophenol), bearing accession numbers AF135269 and AF065166 respectively in GenBank. Over a period of time, it has been observed that our bacterial collection, which is based on physiological and degradative properties, is mostly dominated by genus *Pseudomonas* as identified through bacteriological analysis. In such case, a rapid and specific detection protocol for *Pseudomonas* using its genetic information would be a valuable tool in ecological and diagnostic studies. This was the motivation behind the exercise.

There are programs available for designing polymerase chain reaction (PCR) primer pairs as a means of rapid detection [7,8]. These programs select primer pairs based on the user-defined parameters, such as length, secondary structure of primers, G/C content of primers and amplicons, stability of primers etc. However, they do not provide any information regarding the specificity of the oligonucleotides / patterns; and especially when the list of candidate primer pairs is long, it becomes arduous to select the best primers. Multiple alignment tools are often used to identify specific oligonucleotides [9–12]. These tools when applied to related sequences results into a consensus pattern of conserved and variable sites across the sequences. The basic purpose of multiple alignment is to compare the similarity of sequences and to identify homologous genes from the database. Because they compare the entire sequence, they are sometimes not well suited for identifying short specific oligonucleotide sequences. Keeping this in view, a program HYB*simulator* was developed to design target specific oligonucleotides through computer hybridization simulation (CHS) [13]. The program generates a ProbeSet of candidate oligonucleotides from the target gene sequence. The parameters like salt and oligonucleotide concentrations, melting temperatures are specified to obtain the final ProbeSet meeting these specifications. The CHS is then executed for the ProbeSet against the GenBank database of interest and finally best probes are selected based on the G/C contents, hairpin formation energy and the highest specificity to the database. Recently, another software PRIMROSE was developed to pick specific 16S rRNA probes and PCR primers as ecological tools in the identification and enumeration of bacteria [14]. The program generates oligonucleotides from aligned as well as unaligned sequences. With aligned sequences, the algorithm identifies the consensus sequence and then generates oligonucleotides of specific length from the consensus sequence. A search string is created from each oligonucleotide and database sequences are searched for a match with the generated strings. A string with the number of specific hits exceeding the minimum threshold is considered as important. The program also has an alternative algorithm to deal with unaligned sequences to finally yield the target specific oligonucleotides.

This paper describes an approach, using 16S rDNA sequence data, to select genus specific probes / patterns that could be subsequently used as primers for tracking of bacterium from the environmental niches. The approach is based on the hypothesis that 16S rDNA sequences representing different species of a particular genus, if aligned, then much of the variability across the sequences is due to substitution of bases and hardly by insertion or deletion of bases. The interest is to know whether such substitution prone / mismatched region(s) possesses patterns that are genus specific. And secondly, could there be an alternative way to identify such regions without taking recourse to multiple alignment of complete sequences. One possible way could be to locate some markers that remain preserved across the selected sequences and focus on the region(s) flanked by them; and one such marker could be the repeating patterns of nucleotides. We found that in a set of closely related sequences there are some repeating patterns, which consistently occur across the set; and importantly the separating distance between the repeats is also preserved throughout. Accordingly, the search for mismatched or the substitution prone base positions, was restricted to the sub-sequence enclosed between the highly consistent repeats. The patterns generated from the mismatched region could be tested for their specificity against the standard database. By this, we are ensuring first the variability of patterns across the species of a selected genus, followed by their specificity to the target genus. Thus, the main difference between this approach and the one followed in HYB*simulator* or PRIMROSE is that, in this case the probe / pattern set generation is restricted to only the mismatched region.

The paper details a case study for genus *Pseudomonas* using 16S rDNA sequence data. The most consistent repeating patterns across 16S sequences representing different species of this genus were obtained using a program *Repeat Tuple Search*. A region of fixed length flanked by the repeat markers was obtained. A consensus pattern was obtained for this region using Shanon's entropy measure. A most variable stretch from this region was identified, which provided *Pseudomonas* specific patterns and subsequently a signature pattern. A PCR primer pair was designed considering the conserved and the variable stretch and was validated through a preliminary experiment.

Results

Consistent repeating patterns in Pseudomonas 16S rRNA gene

The training data comprised of fifty different 16S rRNA sequences, representing different species of Pseudomonas retrieved from GenBank (EMBL, Release May 1999). During sampling, partial sequences were ignored and only those having size of around 1.5 kb were considered in the sample. To arrive at the signature pattern, the approach as described in Figure 1 was followed. The sequences were used as input sequences to the program Repeat Tuple Search. The four most consistent repeating patterns of varying sizes that top the frequency distribution with more than 80% consistency across the samples and satisfying the criterion of constant separating distance are shown in Figure 2 (Sequence Accession No. AB013253). The repeat-

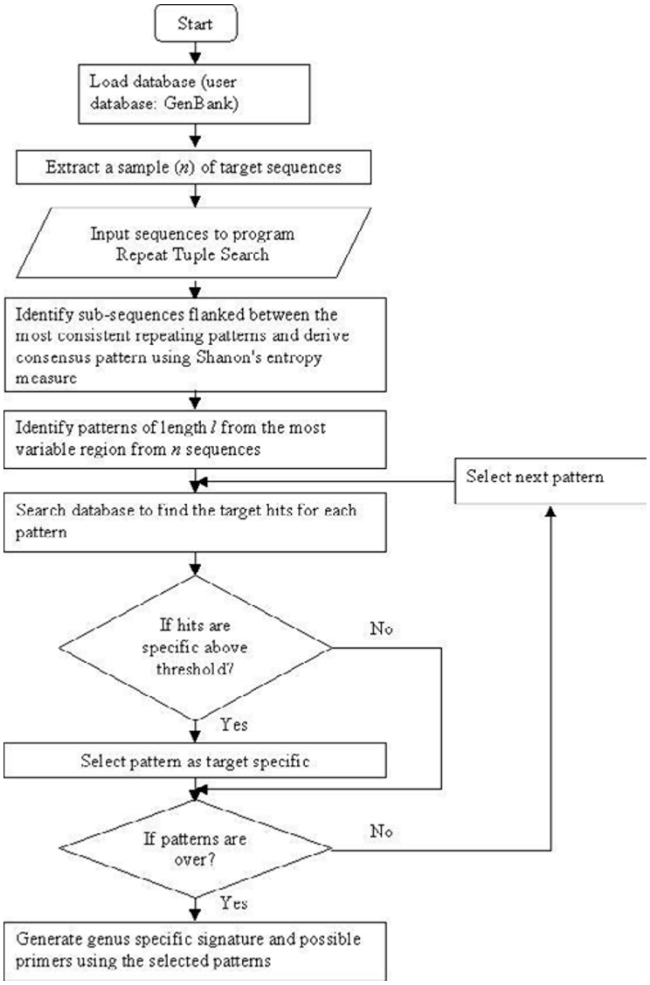
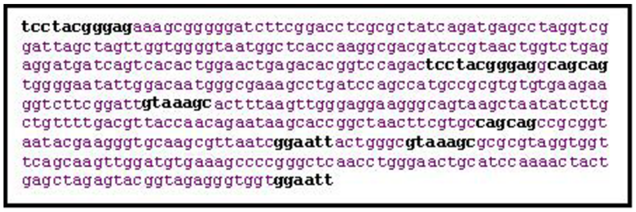

Start
Load database (user database: GenBank)
Extract a sample (n) of target sequences
Input sequences to program Repeat Tuple Search
Identify sub-sequences flanked between the most consistent repeating patterns and derive consensus pattern using Shanon's entropy measure
Identify patterns of length l from the most variable region from n sequences
Select next pattern
Search database to find the target hits for each pattern
If hits are specific above threshold?
No
Yes
Select pattern as target specific
If patterns are over?
No
Yes
Generate genus specific signature and possible primers using the selected patterns

Figure 1
Flow chart summarizing the procedure for identifying target specific patterns



```
tcctacgggagaaagcgggggatcttcggacctcgcgctatcagatgagcctaggtcg
gattagctagttggtggggtaatggctcaccaaggcgacgatccgtaactggtctgag
aggatgatcagtcacactggaactgagacacggtccagactcctacgggaggcagcag
tggggaatattggacaatgggcgaaagcctgatccagccatgccgcgtgtgtgaagaa
ggtcttcggattgtaaagcactttaagttgggaggaagggcagtaagctaatatcttg
ctgttttgacgttaccaacagaataagcaccggctaacttcgtgccagcagccgcggt
aatacgaagggtgcaagcgttaatcggaattactgggcgtaaagcgcgcgtaggtggt
tcagcaagttggatgtgaaagccccgggctcaacctgggaactgcatccaaaactact
gagctagagtacggtagagggtggtggaatt
```

Figure 2
A substitution prone region marked by the four most consistent repeating patterns in 16S rDNA sequence of *Pseudomonas*. Indel free region derived from the selected fifty sample sequences. The most consistent repeating patterns are TCCTACGGGAG, CAGCAG, GTAAAGC, and GGAATT repeating at 156, 173, 142, 116 distances respectively in the sampled sequences.

ing patterns along with their sub-sequences spanned a region of 495 bases in different sequences. It could be seen from the figure that the sub-sequence enclosed between the repeating pattern CAGCAG partly overlaps with that enclosed between the repeating pattern GTAAAGC. Also, a small portion of sub-sequence between GGAATT overlaps with that of GTAAAGC, while the sub-sequence bounded by the repeating pattern TCCTACGGGAG does not overlap with any of these sub-sequences. These repeating patterns are preserved in majority of the sequences and the respective separating distances also remain constant across the sequences resulting into a sub-sequence (region) flanked by fixed markers, which is nearly 1/3 rd length of the 16S sequence. The sub-sequences between the respective repeating patterns from the sampled sequences were aligned using CLUSTAL W program and evaluated for positional conservation using Shanon's entropy.

Entropy analysis

The entropy analysis for the aligned sub-sequences resulted into position specific information content using expression (1) as described in methods. The sub-sequences between the selected four repeating patterns, from fifty sampled 16S sequences were analysed with the results shown graphically in Figure 3A,3B,3C,3D. The alignment in Figure 3A indicate very few base positions with absolute conservation (*IC* = 2). Most of the positions indicated variability with *IC* values ranging between 1.2–1.9 as also indicated by the spikes spread across the length of the sub-sequence. In Figure 3B, the sub-sequence between the repeating pattern CAGCAG shows a good degree of conservation for the first 100 bases, but the remaining 73 bases shows considerable randomness at different positions. A stretch of nearly twenty-five bases between positions 100–125 (with reference to 5'-CAGCAG) shows maximum variability as evident from the deep spikes. The *IC* value at position 124 touched the lowest i.e. 0.2 indicating maximum uncertainty at this position. The pattern in the left of Figure 3C resembles to that of the pattern after base position 75 in Figure 3B, since there is a partial overlap. The sub-sequence flanked by the repeating pattern GGAATT (Figure 3D) is conserved at few positions, while most of the positions indicate high randomness scattered throughout the length of sub-sequence. It is essential to mention here that the analysis of aligned sub-sequences of fixed lengths showed that there are few gapped columns due to insertion or deletion of one or two bases in some sub-sequences. But most of the variation across the species was observed due to substitution.

Thus the conservation analysis for sub-sequences enclosed between the four selected consistent repeats yielded a consensus pattern of conserved and variable positions for the 495 bp region as shown in Figure 4. The

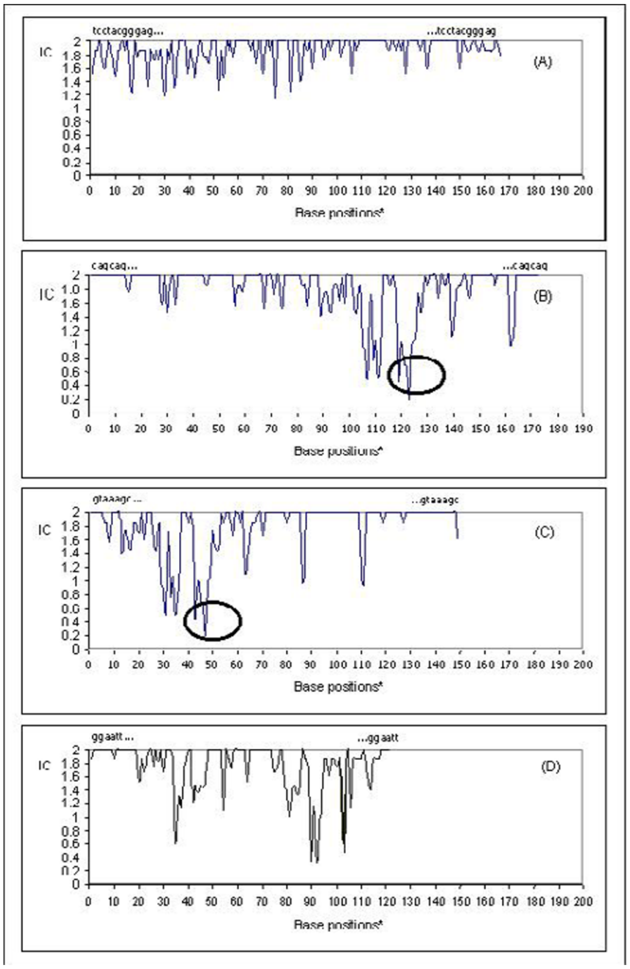
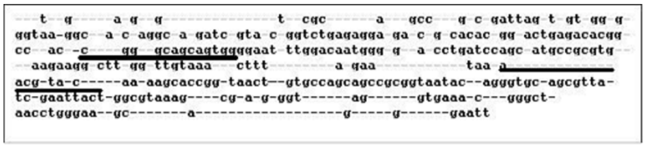


Figure 3
(A-D): The information content (*IC*) measuring the randomness of bases in the columns of alignment. The circles indicate the maximum variable positions after alignment.

region might be having part of the ancestral information as well as some information specific to genus *Pseudomonas*. The 5' and the 3'-ends of this consensus pattern have elongated variable stretches. The alignment of subsequences flanked by the repeating pattern CAGCAG enclosing the stretch of maximum variation, as shown in Figure 3B, was focused for identifying target specific patterns and thereby the primers.

Target specific patterns and PCR primers

A pattern of 24 bases was derived from this stretch from each of the selected fifty sample sequences. Thirty non-duplicating patterns were obtained out of fifty. Each pattern was searched for a match in rRNA database of GenBank using *GeneMan* module of *Lasergene* (*DNAStar Inc.*) software. The number of hits specific to genus *Pseudomonas* was recorded for each pattern with degeneracy allowed at four positions. Amongst all the patterns, nine patterns indicated specificity of nearly 70% to the target genus *Pseudomonas* and are shown in Figure 5. When the composition of bases in these patterns was studied it was observed that the bases in columns indicated by lighter shade are highly preserved not only in species of *Pseudomonas* but also in many other bacteria [data not shown]. Nevertheless the patterns are target specific, implying that there is some more information with these patterns, which is *Pseudomonas* specific and not shared by others. The columns with dark shades precisely reveal this information. The conservation of bases in these columns is observed in nearly 70% of the *Pseudomonas* 16S sequences, unlike in other bacterial genera. This eventually yielded a consensus pattern ACnTnnnTGTnTTGACGTTACCnA, which could be treated as a most representative signature for *Pseudomonas* in the identified region. The nine patterns could be used for designing target specific primers. A sample exercise was done with pattern ACTTTGCTGTTTTGACGTTACCGA, which also showed maximum specificity to genus *Pseudomonas*. The *PrimerSelect* module of *Lasergene* was used to generate primer pair, which yielded a 150 bp product. The reverse primer 5'-TCGGTAACGTCAAAACAGCAAAGT-3' was used as reference for selection of a corresponding forward primer 5'-CTACGGGAGGCAGCAGTGG-3' from the identified region as shown in Figure 4. During the design, it was ensured that the primer has a conserved 3'-base 'A' so as to facilitate the extension of the product if used in the experiment. The forward primer, though indicate relatively conserved base positions, picked many non-specific hits during search confirming that the primer is not just the characteristic of *Pseudomonas*, but is shared by several other genera as well. Likewise, other patterns from this stretch could also be tried out to generate possible *Pseudomonas* specific primer sets for PCR.

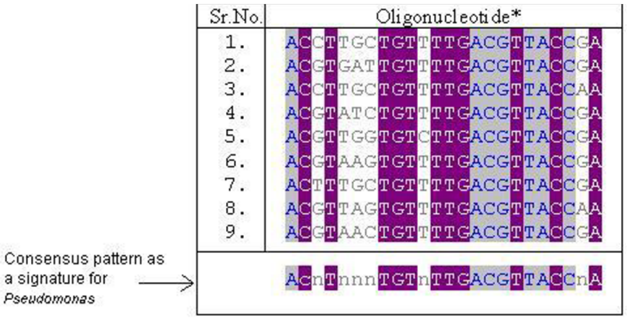


```
---t--g----a-g--g--------------t--cgc-----a---gcc---g-c-gattag-t-gt-gg-g-
ggtaa-ggc--a-c-aggc-a-gatc-gta-c-ggtctgagagga-ga-c-g-cacac-gg-actgagacacgg-
cc--ac--c----gg--gcagcagtggggaat-ttggacaatggg-g--a-cctgatccagc-atgccgcgtg--
--aagaagg-ctt-gg-ttgtaaa---cttt-------a-gaa----------taa-a-----------
acg-ta-c-----aa-aagcaccgg-taact--gtgccagcagccgcggtaatac--agggtgc-agcgtta-
tc-gaattact-ggcgtaaag----cg-a-g-ggt------ag------gtgaaa-c---gggct-
aacctgggaa--gc-------a------------------g-----g------gaatt
```

Figure 4
The consensus pattern obtained for the 495 bp region. Each dash indicates that the *IC* value at that position is less than 1.9, while the bases indicate absolute conservation at the position.

Species level variation within 150 bp PCR product

In order to ensure that the 150 bp sub-sequence bounded by the forward and reverse primer represents diversity



Sr.No.	Oligonucleotide*
1.	ACCTTGCTGTTTTGACGTTACCGA
2.	ACGTGATTGTTTTGACGTTACCGA
3.	ACCTTGCTGTTTTGACGTTACCAA
4.	ACGTATCTGTTTTGACGTTACCGA
5.	ACGTTGGTGTCTTGACGTTACCGA
6.	ACGTAAGTGTTTTGACGTTACCGA
7.	ACTTTGCTGTTTTGACGTTACCGA
8.	ACGTTAGTGTTTTGACGTTACCAA
9.	ACGTAACTGTTTTGACGTTACCGA
	ACnTnnnTGTnTTGACGTTACCnA

Consensus pattern as a signature for *Pseudomonas*

Figure 5

Patterns showing nearly 70% specificity to genus *Pseudomonas* and a consensus pattern as a signature for *Pseudomonas*.

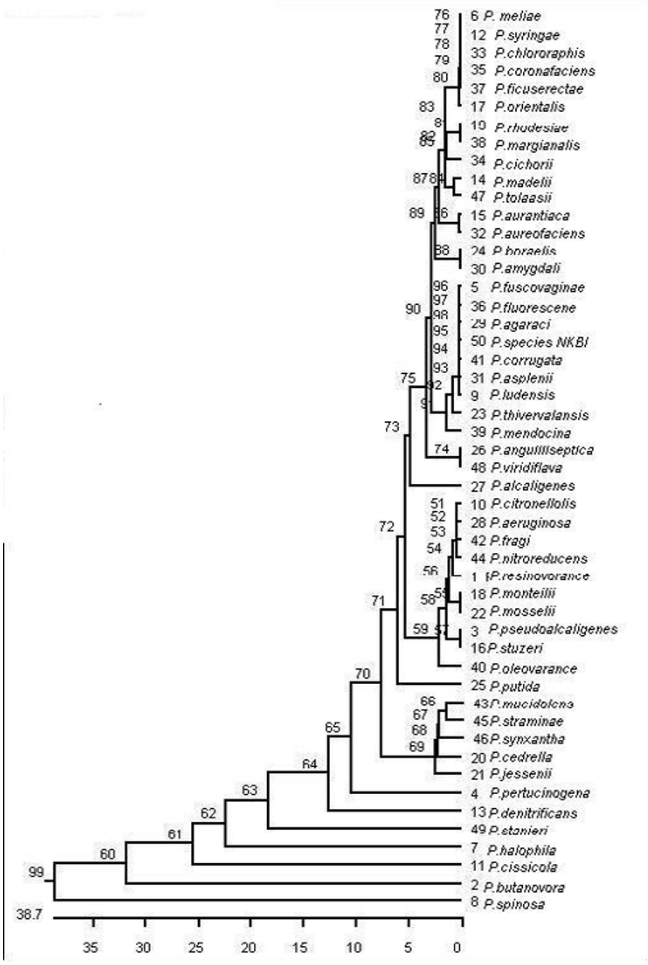


Figure 6

Dendogram showing the species level variation of the selected fifty *Pseudomonas* species based on the 150 bp subsequence.

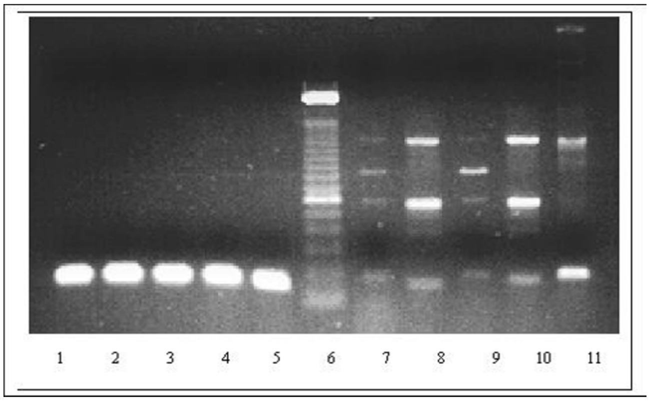


Figure 7

Demonstration of specificity of designed PCR primers for *Pseudomonas*. The following oligonucleotides were designed as forward 5'-CTACGGGAGGCAGCAGTGG-3'and reverse 5'-TCGGTAACGTCAAAACAGCAAAGT-3' for PCR. Lane 1–5: Five different reported *Pseudomonas* strains: SF1 & PH1 (laboratory isolates), Pp2 & CF600 (Received from Dr. V. Shingler, Umea university, Sweden), P. putida (received from Dr. S. Harayama, MBI, Japan); Lane 6: 100 bp ladder; Lane 7: *E. coli*, Lane 8: *Salmonella*, Lane 9: *Vibrio*, Lane 10: *Shigella* Lane 11: *Staphylococcus.* A PCR program of 95°C, 5 min, and 35 cycles of 94°C, 30 s; 62°C, 15 s; and 72°C for 30 s were used to amplify a 150 bp product using total DNA as template. The designed primers were used at 25 pM concentration of each in 50 μl reaction mixture as per the conditions described by the supplier (Perkin Elmer). Lane 1–5 shows the amplification of expected *Pseudomonas* specific product; and lanes 7–11 the non-specific extension of primers with strains negative for the selected conserved region. The quality of the DNA as a template and the control for PCR were checked using the universal eubacterial primers (data not shown).

amongst the samples, cluster analysis was carried out using 150 bp sub-sequences. The sub-sequences were retrieved from the selected fifty *Pseudomonas* sequences and were aligned by CLUSTAL W program, which finally yielded a dendogram. Figure 6 shows the association of various species obtained using *MegaAlign* module of *LaserGene* software. Since, the analysis reasonably supports the species level diversity within the limiting 5' and 3' designed primer pair, a preliminary experimental validation was carried out. Figure 7 shows the optimized amplification reaction using the designed primer set. Six different *Pseudomonas* isolates from various labs were used to derive the DNA template along with the negative controls. The expected 150 bp product with the *Pseudomonas* DNA template ascertained the relative specificity of the primers. The results were further supported by Slot-Blot hybridization analysis as shown in Figure 8 using the independent probes derived from each template and hybridized to the

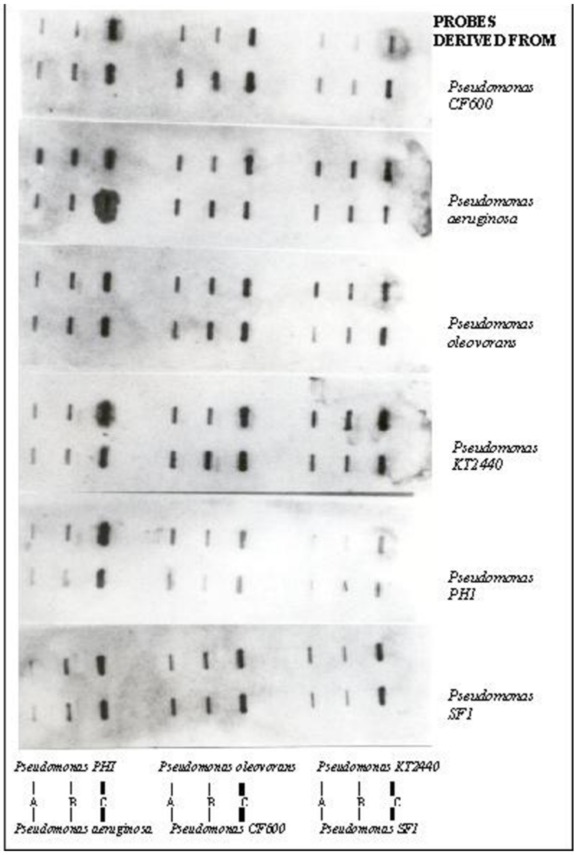


Figure 8
Demonstration of specificity of PCR amplified product for genus *Pseudomonas* by Slot-Blot hybridization data. PCR product was generated using signature primer as shown in Figure 7. Biotinylated probes were generated from PCR amplified 150 bp products using the BioPrime DNA Labeling Kit, Gibco BRL, USA. Six probes were derived using six different *Pseudomonas* strains as mentioned on the right panel of the figure. The same six strains were used as templates in the hybridization as shown in the schematic representation at the bottom of the figure. Each probe was tested against all six strains. The cultures were grown overnight at 30°C and diluted to yield 1000, 5000 and 10,000 cells per 100 μl of sterile double distilled water. This was transferred to a nylon membrane using the Slot-Blot apparatus at positions A, B and C respectively, where A represents 1000 cells, B, 5000 cells and C represents 10,000 cells. The spotted membranes were immersed in SDS/BSA hybridization buffer (0.3 ml/cm^2 of membrane). Hybridization was carried out at 60°C in a hybridization oven, Amersham Pharmacia Biotech Ltd., UK. Hybridization signals were detected using the Photogene Nucleic Acid Detection System Version-2 (Gibco-BRL, USA), as per the manufacturer's instructions with high stringency washing. Chemiluminiscent signals were quenched on Kodak: X-omat film after exposure of 30 min. The negative controls used were *E. coli*, *Salmonella*, *Vibrio*, *Shigella* and *Staphylococcus* as used in Figure 7 to show the specificity of probe. There were no hybridization signals observed in any of the negative controls (data not shown) even after overnight exposure.

selected all the strains of *Pseudomonas*. The same probes when used against negative controls as used in Figure 7, there were no signal observed even after overnight exposure.

Discussion

Amongst several approaches to determine target specific patterns, we have recently proposed one based on the dinucleotide compositions that discriminate a group of selected bacteria from each other [15]. The selected dinucleotides could be used to generate patterns whose specificity could be tested through search analysis against the standard databases.

However, in this case, repeating patterns that are conserved across different sequences of *Pseudomonas*, have been used to locate a mismatched region and eventually for selection of genus specific pattern / signature and primers. The identified four repeats exhibit a pattern of repeating elements, which is quite dominant across the *Pseudomonas* 16S sequences, as shown in Figure 2. In other studies, we have identified repeating elements in 16S sequences of fifty different bacterial groups http://www.ims.nus.edu.sg/Programs/genome, but the occurrence of these four repeats, together, was not observed in other groups except *Pseudomonas*. This suggests that in other genera, either these patterns have undergone substitutions at one or more corresponding base positions thereby not constituting repeats or might be that these four conserved repeating patterns in *Pseudomonas* is the result of evolution. Some recent studies have suggested the evolutionary significance of repeat elements, following the discovery of similar repeat units among the species of vertebrate [16,17]. Likewise, in *Pseudomonas* also, the identified consistent repeats might have some relevance, which could be explored.

For large size DNA fragment analysis, the approach based on tandem repeats provides a reference through which a sequence could be characterized. Using tandem repeats as reference, PCR primers could be designed to differentiate heterologous or even homologous DNA fragments [18]. In this study, a similar strategy was adopted to design PCR primers, but using simple repeating patterns as reference. The most consistent repeating patterns were used as markers to define a sub-sequence of fixed length in selected sequences. As regards the difference between this approach and the one proposed in PRIMROSE, in the later the program generates large number of possible patterns using the consensus pattern for a whole sequence and then determines the target specific patterns based on the statistics of number of hits. Similarly, in HYB*simulator* also, the program generates probe set of specified length across the length of the input sequence. Contrary to this, in the present approach, the probe set is restricted to only the substitution prone or the mismatched region of the alignment. The determination of specificity of pattern is similar to that of PRIMROSE and based on the statistics of target and non-target hits, and does not consider any thermodynamic properties of probes as in the case of HYB*simulator*. Moreover, the approach depicts part of the species level diversity finally yielding a signature pattern for the target genus, which is an interesting observation of this study. The signatures if ascertained for their specificity could be used for typing the bacteria in different habitats. Here, the signature was used to design a tracking protocol for genus *Pseudomonas*, which was also validated on a preliminary scale through experiment. However, a rigorous validation work plan would be required with more DNA templates. The exercise should also include significant number of unknown isolates from environmental niches. The DNA derived from these isolates could be tested for 150 bp product and further support by ARDRA of 16s rDNA using band-sharing index. Alternatively, the DNA templates yielding 150 bp product could be randomly picked to derive complete 16S rDNA sequences to ascertain the specificity of the approach to genus *Pseudomonas*.

Conclusion

We had an assumption that the species level differences could be captured by nucleotide patterns in the mismatched or substitution prone sites in sequences belonging to a particular genus. This has been verified through a sample experiment by analysing a small region of 16S rDNA bounded by repeating patterns. There could be some other position specific markers such as non-duplicating patterns; however, the constant distance criterion between such patterns across the sequences needs to be ascertained as in case of repeating patterns. Although, the approach was implemented successfully for *Pseudomonas*, it may have limitations while applying in other genera. For instance, in some cases, the sub-sequences flanked by consistent repeats might indicate lesser variability across the length of sub-sequence or the variability might be uniform throughout its length. In that case identifying a stretch to select patterns would be difficult. Further, it might also happen that the patterns belonging to most variable stretch does not produce sufficient hits to be considered as target specific. We would like to state here that the strength of our assumption and the approach depends much on its applicability in other bacterial genera and recommend that it may be investigated further. We are trying to use this approach to derive signatures and primers for different genera of importance in bioremediation. If it works for other genera and provides genus specific signatures then it could be one of the possible means of designing micro arrays for studying population dynamics in bioremediation protocols or similar kind of applications.

Methods

The procedure for selecting target specific patterns has been summarized through flow chart as shown in Figure 1 and is as follows:

Consistent repeating patterns in related sequences

A repeat pattern is a sub-string of nucleotides in a sequence *S*, which occurs more than once in *S*. The origin, evolution and distribution of repetitive elements in genome databases have been the subject of intense study, both experimentally and computationally. Some algorithms to find repeats in a sequence could be found in [19–21]. Amongst these, the recently developed *REPuter* has been found to be efficient and provides exhaustive repeats, even in large genome sequences. The algorithm uses a compact implementation of suffix trees to locate exact repeats in linear space and time for sequences.

Although the above programs provides a list of repeats in a sequence, they do not have an option to automatically provide repeating patterns, which are conserved across the input set of sequences, by simultaneously considering the spacings criterion. In fact, processing data on repeats from different input sequences could be a small extension to many of these programs. We have developed a program *Repeat Tuple Search*, which has this additional feature to determine the consistent repeating patterns across the set of sequences. The program could be downloaded from site http://www.ebi.ac.uk/~lijnzaad/RepeatTupleSearch and used to handle small gene sequences up to size 2 kb. It accepts sequences one-by-one in a simple text format and stores data on repeating patterns for each input sequence. The program has two basic components – the first determines the repeating patterns of lengths more than six (default setting) along with the spacing between the repeats in each input sequence. The search for repeating patterns is exhaustive without asking for any input conditions from the user. The second component processes the collective data to get the number of sequences in which different repeating patterns make their appearances, considering the constant spacing criterion. The patterns with high frequency of occurrence are considered as the most consistent repeating patterns.

Entropy-based conservation analysis

Once the consistent repeating patterns are determined, the sub-sequences enclosed between them could be studied for the degree of conservation across the set of input sequences. The sub-sequences enclosed between a particular repeat from different sequences could be aligned using CLUSTAL W program to arrive at the positional conservation. The *Shanon's entropy* could be used as a measure of positional conservation. It gives the average uncertainty of an outcome at different positions using expression

$$S(m_i) = -\sum_a p_{ia} \times \log_2 p_{ia} \qquad ...(1)$$

where, m_i stands for the i^{th} position and p_{ia} is the probability of residue a in i^{th} position with reference to 5'-end tuple and could be estimated using the maximum likelihood estimate

$$p_{ia} = \frac{c_{ia}}{\sum_{a'} c_{ia'}} \qquad ...(2)$$

where c_{ia} is the total number of a's in i^{th} position, while the denominator in (2) gives the total number of bases at that position. The logarithm in expression (1) has base 2, hence the unit of entropy is a *bit*. The information content could be used, which gives the reduction in uncertainty after some 'knowledge' has been received. In other words, information content at a particular position is the difference between the entropy before and after the knowledge, and is given by

$IC(m_i) = S_{before} - S_{after}$

Without any bias, the assumption could be that the bases in a particular column are random i.e. $p_a = 0.25$, giving $S_{before} = 2$ bits. On observing the occurrence of bases in the columns of alignment, S_{after} could be obtained and thus the difference gives the information content of the column. Such information content could be obtained at different positions of the aligned sub-sequences, thereby yielding conserved and variable base positions [22].

The most variable stretch of nucleotide positions could be identified and the patterns of some fixed length '*l*' spanning the variable stretch from the sampled sequences could be selected. The specificity of each selected pattern could be tested against the rRNA database. If the specific hits produced by a pattern exceed some predetermined threshold, select it as target specific, otherwise, the search is aborted and the next pattern is picked. Once the search for all the patterns is over, the selected target specific pattern could be used to generate a signature pattern, which would be specific to the target locus. Also, each selected pattern could be tested for its possibility as a primer using any of the primer selection programs to facilitate rapid detection of the target bacterium in environmental samples.

Author's Contribution

The authors have equally contributed to the manuscript.

HJ Purohit: For developing the concept, retrieving the sequences, participated in sequence alignment using Lasergene software, developing dendogram, designing PCR primers followed by the validation through molecular experiment.

DV Raje: Developed the algorithm to determine consistent repeating patterns in a set of homologous sequences, entropy analysis to finally arrive at the signature pattern for genus *Pseudomonas* through search analysis.

A. Kapley: Experimental validation with Slot-Blot hybridization using different *Pseudomonas* strains.

References

1. Atuanya EI, Purohit HJ and Chakrabarti T **Anaerobic and aerobic biodegradation ofchlorophenols using UASB and ASG bioreactors** *World Journal of Microbiology & Biotechnology* 2000, **16:**95-98
2. Chhatre SA, Purohit HJ, Shanker R, Chakrabarti T and Khanna P **Bacterial consortia for crude oil spill remediation** *Water Science and Technology* 1996, **34:**187-193
3. Kapley A, Purohit HJ, Chhatre S, Shanker R, Chakrabarti T and Khanna P **Osmotolerance and hydrocarbon degradation by genetically engineered bacterial consortium** *Bio-Resource Technology* 1999, **67:**241-245
4. Kapley A, Lampel K and Purohit HJ **Thermocycling steps and optimization of multiplex PCR** *Biotechnology Letters* 2000, **22:**1913-1918
5. Kapley A and Purohit HJ **Tracking of phenol degrading genotype** *Environmental Science and Pollution Research* 2001, **8:**89-90
6. Kutty R, Purohit HJ and Khanna P **Isolation and Characterization of a *Pseudomonas sp.* strain PH1 utilizing meta-aminophenol** *Canadian Journal of Microbiology* 2000, **46:**211-217
7. Lowe T, Sherfkin J, Yang SQ and Dieffenbach CW **A computer program for selection of oligonucleotide primers for polymerase chain reaction** *Nucleic Acids Research* 1990, **18:**1757-1761
8. Mitsuhashi M, Cooper A, Ogura M, Shinagawa T, Yano K and Hosokawa T **Oligonucleotide probe design – a new approach** *Nature* 1994, **367:**759-761
9. Feng DF and Doolittle RF **Progressive sequence alignment as a prerequisite to correct phylogenetic trees** *Journal of Molecular Evolution* 1987, **25:**351-360
10. Lipman DJ, Altschul SF and Kececioglu JD **A tool for multiple sequence alignment** *Proceedings of National Academy of Science* 1989, **86:**4412-4415
11. Thompson JD, Higgins DG and Gibson TJ **CulstalW: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice** *Nucleic Acids Research* 1994, **22:**4673-4680
12. Thompson JD, Plewniak F and Poch O **A comprehensive comparison of multiple sequence alignment programs** *Nucleic Acids Research* 1999, **27:**2682-2690
13. Hyndman D, Cooper A, Pruzinsky S, Coad D and Mitsuhashi M **Software to determine optimal oligonucleotide sequences based on hybridization simulation data** *BioTechniques* 1996, **20:**1090-1096
14. Ashelford KE, Weightman AJ and Fry JC **PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database** *Nucleic Acids Research* 2002, **30(15):**3481-3489
15. Raje DV, Purohit HJ and Singh RN **Distinguishing features of 16S rDNA gene for five dominating bacterial genus observed in bioremediation** *Journal of Computational Biology* 2002, **9(6):**819-829
16. Fumigalli L, Taberlet P, Favre L and Hauser J **Origin and evolution of homologous repeated sequences in mitochondrial DNA control region of shrews** *Molecular Biology and Evolution* 1996, **13:**31-46
17. Wilkinson GS, Mayer F, Kerth G and Petri B **Evolution of repeated sequence arrays in evening bat D-loop mtDNA** *Genetics* 1997, **128:**607-617

18. Liu BH **Statistical Genomics: Linkage, Mapping and QTL analysis** *CRC Press LLC, Florida* 1998,
19. Rigoutsos I and Floratos A **Combinatorial Pattern Discovery in Biological Sequences: the TEIRESIAS Algorithm** *Bioinformatics* 1998, **14:**55-67
20. Kurtz S and Schleiermacher C **REPuter: fast computation of maximal repeats in complete genomes** *Bioinformatics* 1999, **15:**426-427
21. Kurtz S, Choudhari J, Schleiermacher C, Stoye J and Giegerich R **REPuter: the manifold applications of repeat analysis on a genome scale** *Nucleic Acids Research* 2001, **29:**4633-4642
22. Durbin R, Eddy S, Krough A and Mitchison G **Biological Sequence Analysis: Probabilistic models for proteins and nucleic acids** *Cambridge University Press* 1998,