Pyrosequencing with di-base addition for the identification of pathogenic bacteria

Dan Pu¹, and Kunxian Shu ¹, *

¹ Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract. We proposed a novel method for the identification of pathogenic bacteria, in which a partial amplicon of the molecular markers was targeted by using pyrosequencing with di-base addition (PDBA). PDBA was conducted by synchronously adding di-base instead of one base into a reaction, and a set of highly sequence-specific encodings containing the type and the number of incorporated nucleotide(s) (peak height) were obtained. By comparing the encoding sizes of each isolate and the number of incorporated nucleotide(s) in each cycle, moving from first to last, various kinds of bacteria could be unambiguously identified. To verify its feasibility, we simulated PDBA results from thirteen isolates of Mycobacterium species and compared their encoding sizes and the number of incorporated nucleotide(s) in each cycle with those predicted by a homemade software. The thirteen isolates were successfully differentiated. We also targeted partial RNase P RNA gene (rnpB) of cultured M. paratuberculosis and M. celatum to differentiate the two isolates. By comparing the encoding size of each isolate and the number of incorporated nucleotide(s) in each cycle, the two Mycobacterium isolates were successfully differentiated. In conclusion, PDBA enabled to reliably identify pathogenic bacteria.

1 Introduction

It is important to identify pathogenic bacteria in the field of microbiological diagnostics and thus the demand for simple, sensitive, specific, and robust tests is increasing. For the identification of pathogenic bacteria, both biochemical testing and molecular methods are developed. However, biochemical testing may yield inconsistent or inconclusive results ([1]). Molecular methods provide microbiologists with additional tools that may supplement biochemical testing for bacterial pathogen identification ([2]). Among the molecular methods, DNA sequencing technologies, such as Sanger sequencing, pyrosequencing and next-generation sequencing (NGS), offer attractive strategies for the identification of pathogenic bacteria from clinical specimens and it is regarded as a straight way to differentiate pathogenic bacteria. Sanger sequencing is low-cost and high-accuracy. But its low throughput limits it to be used to simultaneously identify large number of pathogenic bacteria. Pyrosequencing is a sequencing-by-synthesis based method, which has been widely used to detect pathogenic bacteria by targeting molecular markers, such as rnpB ([3]), rpoB ([4]), 16s rRNA ([5]), gyrB ([6]). Traditional pyrosequencing targets species-specific nucleotide sequences by adding one nucleotide into each reaction at a time so that the type and the number of incorporated nucleotide(s) can be determined. The isolates are identified from sequence alignment of a large sequence dataset, such as GenBank and the partners of the International Nucleotide Sequence Database Collaboration ([3]). However, it is difficult to obtain concrete sequences by using traditional pyrosequencing when the abundance of analyzed isolates is relatively low and species-specific nucleotide sequences is longer than 60 bp ([3, 7]). Furthermore, the limitations in resolution of closely related species hamper it as a routine tool for identify pathogenic bacteria. NGS is another sequencing technique enables to identify pathogenic bacteria. It is rapid and provides confident identity of any bacteria present in a sample. However, NGS is so sensitive that it can detect bacteria present in very small numbers, below the infection threshold, resulting in unnecessary antibiotic therapy ([8]). Therefore, it is necessary to explore an additional method for the identification of pathogenic bacteria. We have proposed a novel sequencing method in which a template is cyclically sequenced twice in two parallel sequencing runs with different combinations of di-base addition (AG/CT, AT/CG, or AC/GT) and two sets of encodings are obtained ([9]). The two sets of encodings allow for the bases to be sequentially decoded, moving from first to last, in a deterministic manner. This approach is compatible with most of the SBS-based sequencing platforms, such as traditional pyrosequencing, Roach 454, Ion torrent PGM ([9]). When the novel sequencing technique is combined with traditional pyrosequencing, we call it pyrosequencing with di-base addition (PDBA). Due to the addition of di-base into the reaction, PDBA enables to increase signal intensity (peak height) in each sequencing run.
run and to reduce the number of the sequencing cycles (A sequencing cycle represented a sequencing reaction) required to read though the target sequences. Thus, when specific target sequences are interrogated by PDBA, encodings containing the peak height and the number of sequencing cycles (encoding size) are sequence-specific. By comparing the encoding sizes and the peak height of species-specific nucleotide sequences with those predicted by a homemade software, the isolates can be unambiguously differentiated.

Here, we applied PDBA to differentiate pathogenic bacteria. The target for PDBA analysis was the rnpB gene coding for the RNA subunit of RNase P. The gene was present in all bacteria and has both conserved and highly variable sequence regions. It has shown to be an excellent target for differentiation of bacterial species of the diverse genera ([3]). In the present study, take Mycobacterium species as an example, partial rnpB gene of Mycobacterium species was interrogated by PDBA. To verify the feasibility of this novel method, we first simulated PDBA results from 13 isolates of Mycobacterium species and then compared the encodings of each isolate with those predicted by use of the homemade software. In addition, to discriminate two strains of Mycobacterium strains, M. paratuberculosis and M. celatum, a portion of the rnpB genes of the two isolates that contained a hypervariable region were also targeted by PDBA.

2 Materials and methods

2.1 DNA preparation

DNA was extracted from two cultured Mycobacterial strains by using QIAamp DNA minikit (QIAGEN, Hilden, Germany) according to the manufacturer’s instructions. The extracted DNA pellets were stored at -20°C for use.

2.2 PCR

PCR assays targeted partial rnpB gene that contained a hypervariable region was used for the sequence-based identification of Mycobacteria. A sequencing primer was designed for the most effective sequencing of the hypervariable region. The templates harboring partial rnpB gene were amplified using the following primers: Forward, 5’- CGGATGAGTTGGCTGGGCGGCCGCGG-3’; Reverse: 5’- biotin- GCTCTTACCGCACCGTTTCACC -3’. Primers were synthesized by Shenzhen Huada Gene Technology Co. LTD. (Shenzhen, China). Reactions for PCR were performed using 25 ng of template DNA in a total volume of 25 μL containing a final concentration of 1× HotStar Taq DNA polymerase Buffer, 1.6 mM of additional MgCl2, 200 μM of dNTPs, 200 nM of each primer and 2 U of HotStar TaqDNA polymerase. The PCR reaction consisted of the following: an initial denaturing at 98°C for 30 s, 45 cycles of 98°C for 10 s, 56°C for 20 s, and 72°C for 15 s, followed by extension at 72°C for 5 min. PCR products were immediately stored at -20°C. Electrophoresis using 2% agarose gels was used to verify PCR products.

2.3 Conventional pyrosequencing

Twenty microliters of PCR products were purified and rendered single-stranded on a PyroMark Q96 Vacuum Workstation (Qiagen, Courtaboeuf, France) following the manufacturer’s instructions. The single-stranded templates and 1 μM sequencing primer (5’ - TGCGTCGGCCGCGCGGG- 3’) were added into a 96-well filter plate and then annealed by incubating in the presence of annealing buffer for 2 min at 80°C. After annealing, the mixture was cooled to room temperature for use. Pyrosequencing analysis was carried out on an automated PSQ 96MA system (Biotage AB, Uppsala, Sweden) by using the SQA PyroMark Gold Q96 Reagents (1×96), containing enzyme, substrate, and nucleotides. The reaction cascaded primarily by cyclic supplying of the C/T/G/A into the growing DNA chains. The pattern of emitted light in relation to the type and number of nucleotides incorporated was subsequently illustrated on a pyrogram. For reproducibility, each sample was also analysed in triplicate.

2.4 Pyrosequencing with di-base addition (PDBA)

According to our previous work, the cartridge contains a total of six reagent vials. Substrates and enzymes are separately added into a reagent vial, di-bases AG, CT, AC, GT, and AT, and CG are selectively added into the remaining four vials ([7]). Compared to conventional pyrosequencing, PDBA was carried out by adding one di-base instead of one base (A, G, C, and T) into the reaction at a time. Here, di-bases AC and GT were separately placed into each reagent vial and cyclically added into the reactions. After interrogating, a set of encodings which contained the number of incorporated base(s) and the type of di-base was sequentially obtained by translating the peaks in the pyrograms.

3 Results and discussion

3.1 The principle for identification of pathogenic bacteria using PDBA

We have proposed a real-time decoding sequencing strategy in which a template is cyclically interrogated in two parallel sequencing runs with two orders of di-base dispensation. The two orders of di-base dispensation are randomly chosen from the three di-base dispensation orders AG/CT, AT/CG and AC/GT. After two parallel sequencing runs, two sets of encodings containing the type and the number of incorporated nucleotide(s) are obtained. By decoding the two sets of encodings obtained, moving from first to last, in a deterministic manner, the queried sequence allows to be sequentially determined ([10]). Interestingly, for each set of
To verify the feasibility of PDBA for the identification of pathogenic bacteria, partial rnpB gene of thirteen Mycobacterial isolates (M. celatum, M. marinum, M. gordonae, M. leprae, M. kansasii, M. gastri, M. malmoense, M. xenopi, M. intracellular, M. fortuitum, M. smegmatis, M. paratuberculosis, M. celatum, M. fortuitum, and M. smegmatis) were downloaded from Genbank and their encodings obtained by PDBA were predicted with the homemade software. As for each isolate, we predicted PDBA results with three kinds of di-base dispensation order (AG/CT, AC/GT and AT/CG) by use of the software, and a total of thirty-nine sets of encodings was obtained. We gave up the orders of di-base dispensation which different strains would obtain the same set of encodings, from the first to the last cycle, all the isolates could be identified.

3.2 The proof-principle experiment

To verify the feasibility of PDBA for the identification of pathogenic bacteria, partial rnpB gene of thirteen Mycobacterial isolates (M. celatum, M. marinum, M. gordonae, M. leprae, M. kansasii, M. gastri, M. malmoense, M. xenopi, M. intracellular, M. fortuitum, M. smegmatis, M. paratuberculosis, M. celatum, M. fortuitum, and M. smegmatis) were downloaded from Genbank and their encodings obtained by PDBA predicted with the homemade software. As for each isolate, we predicted PDBA results with three kinds of di-base dispensation order (AG/CT, AC/GT and AT/CG) by use of the software, and a total of thirty-nine sets of encodings was obtained. We gave up the orders of di-base dispensation which different strains would obtain the same set of encodings, from the first to the last cycle, all the isolates could be identified.
Therefore, the four *Mycobacterial* isolates were successfully discriminated.

![Simulated pyrograms](https://example.com/pyrograms)

Fig 1. Simulated pyrograms from PDBA with di-base dispensation AT/CG for the identification of thirteen *Mycobacterium* isolates (*M. celatum*, *M. marinum*, *M. gordonae*, *M. leprae*, *M. kansasii*, *M. gastri*, *M. malmoense*, *M. xenopi*, *M. intracellular*, *M. paratuberculosis*, *M. celatum*, *M. fortuitum*, and *M. smegmatis*). (A)-(I) were simulated PDBA pyrograms from the thirteen *Mycobacterium* isolates by use of di-base dispensation AT/CG. The encoding of each reaction was shown upon the flowgrams.

The isolates with the same encoding size were grouped together and highlighted in gray in table 1. There were three groups with the encoding sizes of 20, 16, and 8, respectively. The simulated pyrograms from PDBA of the three group isolates were illustrated in Fig. 1. For the first group, each isolate required twenty sequencing cycles to read through the partial *rnpB* sequence (Fig. 1A-C). That meant the encoding size of each isolate in Fig. 1A-C was 20. However, for the three isolates, the number of incorporated nucleotides in some sequencing cycles might be different from each other. By comparing encodings from the pyrograms with that predicted from the software in each cycle, from the first to the last, all the isolates could be identified. For example, the number of incorporated nucleotides in the first cycle was 1, 4, and 1, respectively. When the three sets of encodings were compared with those of expectant encodings from software, *M. gordonae* incorporated 4 nucleotides in the first cycle during the pyrosequencing reaction. Thus, Fig. C represented pyrogram from *M. gordonae*. In Fig. 1A and B, the number of incorporated nucleotides in the first and second cycle were the same. However, the number of incorporated nucleotides in the third cycle was 4 and 2, respectively. When the two sets of encodings were compared with those of expectant encodings from software, *M. marinum* and *M. leprae* incorporated 4 and 2 nucleotides in the third cycle. Thus, Fig. 1A and B were pyrosequencing results from *M. marinum* and *M. leprae*. That was, by using only three pyrosequencing cycles, three *Mycobacterial* isolates were successfully discriminated.

The remaining two groups could be detected by analogy. For the second group, the encoding size of each isolate was 16 (Fig. 1D and E). The number of incorporated nucleotides in the first and second cycles were the same whereas the number of incorporated

---

*Fig 1.* Simulated pyrograms from PDBA with di-base dispensation AT/CG for the identification of thirteen *Mycobacterium* isolates (*M. celatum*, *M. marinum*, *M. gordonae*, *M. leprae*, *M. kansasii*, *M. gastri*, *M. malmoense*, *M. xenopi*, *M. intracellular*, *M. paratuberculosis*, *M. celatum*, *M. fortuitum*, and *M. smegmatis*). (A)-(I) were simulated PDBA pyrograms from the thirteen *Mycobacterium* isolates by use of di-base dispensation AT/CG. The encoding of each reaction was shown upon the flowgrams.
nucleotides in the third cycle was 2 and 4, respectively. When the two sets of encodings were compared with those of expectant encodings from software, M. gastri and M. malmoense incorporated 2 and 4 nucleotides in

Fig 2. Pyrograms from the identification of M. paratuberculosis and M. celatum by use of PDBA and conventional pyrosequencing, respectively. (A) and (B) were PDBA pyrograms from M. paratuberculosis and M. celatum. PDBA was performed by cyclically adding di-bases AC and GT into sequencing reactions. A and T located at the horizontal axis represented di-bases AC and GT, respectively. The encoding of each reaction was shown upon the flowgrams. Arrows indicated the first encodings different between the two isolates. (C) and (D) were pyrograms from the identification of M. paratuberculosis and M. celatum by use of traditional pyrosequencing. Nucleotides were flowed by adding in the order of C/T/G/A. the third cycle. Thus, Fig. 1D and E were PDBA results from M. gastri and M. malmoense. Similarly, Fig. 1F-I were PDBA results from M. paratuberculosis, M. celatum, M. fortuitum and M. smegmatis, respectively. All in all, by comparing each encoding with that predicted by software, from the first to the last, all the isolates could be identified by up to five cycles.

3.3 PDBA for the identification of M. paratuberculosis and M. celatum

The rnpB gene consists of approximately 400 nucleotides and has some regions to vary greatly between species in both nucleotide composition and size, whereas others are conserved for retained enzymatic activity ([16]). The hyperconservative and hypervariable are useful for identifying pathogenic bacteria. In this study, the PDBA results from partial rnpB gene of M. paratuberculosis and M. celatum by use of three sets of di-base dispensation order AG/CT, AC/GT, and AT/CG were predicted by the software (Table 2). As shown, each set of the encodings enabled to clearly discriminate the two isolates, thus AC/GT was randomly chosen for use. Simultaneously, DNA from M. paratuberculosis and M. celatum was extracted and amplicons containing partial sequences of rnpB gene were obtained by PCR. The amplicons were then analyzed by using PDBA with di-base dispensation order AC/GT (Fig. 2). The sequencing results were translated into a set of encodings and each encoding upon the peak was consistent with what we predicted previously. As could be seen from Fig. 2A and B, the two queried isolates required a total of 14 cycles to read through the targeted sequences. However, the number of incorporated nucleotides (peak height) in some cycles were different from each other. From the first to the third cycles, both M. paratuberculosis and M. celatum incorporated one nucleotide during the pyrosequencing reactions. However, in the fourth cycle, isolate in Fig. 2A incorporated 2 nucleotides during the pyrosequencing reaction while isolate in Fig. 2B incorporated 4 nucleotides. When the two sets of encodings were compared with those of the predicted encodings from software, M. paratuberculosis and M. celatum incorporated 2 and 4 nucleotides in the third cycle during the pyrosequencing reaction. Thus, Fig. 2A and B were PDBA results from M. paratuberculosis and M. celatum, respectively. Therefore, by using only 4 cycles, the two Mycobacterial isolates, M. paratuberculosis and M. celatum, were clearly discriminated.

We also targeted the partial rnpB gene sequences from M. paratuberculosis and M. celatum via traditional pyrosequencing (Fig. 2C and D). Conventional pyrosequencing enabled to directly show the type and the number of incorporated nucleotides in each cycle in the pyrograms. As shown, to read through the partial rnpB gene sequences of the two Mycobacterial isolates, conventional pyrosequencing applied 35 cycles while PDBA applied 14 cycles. Furthermore, from the first to the nineth cycle, the peak patterns of the two isolates were the same, but they were not identical until the 10th cycle, which meant conventional pyrosequencing required at least 10 sequencing cycles to discriminate the two isolates. That was, when a random space-specific sequence was analyzed by both conventional pyrosequencing and PDBA, respectively, conventional pyrosequencing approximately needed more cycles to read through it. In addition, due to simultaneous addition of two nucleotides at a time, the signal intensities from PDBA were relatively higher than that from traditional pyrosequencing. This was useful for identifying pathogenic bacteria with lower concentration. When the concentration of isolates was lower, the signal intensity would be lower and lower with the extension of sequencing primers, resulting in more and more false signals. When two nucleotides were simultaneously added into the reaction cycles, the signal
intensity in each cycle was amplified, making low-abundance sequences detectable. Notably, strong background could be seen in Table 2.

| The order of di-base dispensation | M. paratuberculosis | M. celatum |
|----------------------------------|---------------------|------------|
| AG/CT                            | AG\textsuperscript{2}CT\textsuperscript{1}AG/CT\textsuperscript{2}AG\textsuperscript{1}CT\textsuperscript{1}AG\textsuperscript{4}CT\textsuperscript{3}AG\textsuperscript{1} | AG\textsuperscript{2}CT\textsuperscript{1}AG\textsuperscript{2}CT\textsuperscript{1}AG\textsuperscript{1}CT\textsuperscript{1}AG\textsuperscript{1} |
| AC/GT                            | GT\textsuperscript{1}AC\textsuperscript{1}GT\textsuperscript{1}AC\textsuperscript{1}GT\textsuperscript{1}AC\textsuperscript{1}GT\textsuperscript{1}AC\textsuperscript{1} | GT\textsuperscript{1}AC\textsuperscript{1}GT\textsuperscript{1}AC\textsuperscript{1}GT\textsuperscript{1}AC\textsuperscript{1}GT\textsuperscript{1}AC\textsuperscript{1} |
| AT/CG                            | CG\textsuperscript{1}AT\textsuperscript{1}CG\textsuperscript{1}AT\textsuperscript{1}CG\textsuperscript{1}AT\textsuperscript{1}CG\textsuperscript{1}AT\textsuperscript{1} | CG\textsuperscript{1}AT\textsuperscript{1}CG\textsuperscript{1}AT\textsuperscript{1}CG\textsuperscript{1}AT\textsuperscript{1} |

Fig. 2C and D. It might result from the low concentration of single-strand PCR products. However, when those amplicons were analyzed by PDBA, clear peaks with low background and high quality were observed (Fig. 2A and B).

An inherent problem of PDBA, also seen in conventional pyrosequencing, was the imperfect resolution of homopolymers, leading to a decreased reproducibility of the pyrosequencing assay used for microbiology identification. However, this problem might be avoided by choosing a proper order of di-base dispensation for PDBA.

4 Conclusions

We have described a novel method entitled pyrosequencing with di-base addition (PDBA) for identification of pathogenic bacteria which was conducted by the addition of di-bases into sequencing reactions. Sequencing results were translated into a set of encodings containing the type and number of incorporated nucleotide(s). The number of incorporated nucleotide(s) in each cycle and the encoding size of each isolate were highly sequence-specific. Thus, by comparing the encoding size of each isolate and the number of incorporated nucleotide(s) in each cycle, moving from first to last, various kinds of bacteria could be unambiguously identified. We also have simulated PDBA results from thirteen isolates of Mycobacterium species. By comparing encoding sizes of each isolate and the number of incorporated nucleotide(s) in each cycle, the thirteen Mycobacterium isolates were successfully differentiated.

Acknowledgement

The work was funded by the National Science Foundation of China [61801071], the Basic Research and Frontier Exploration Project of Chongqing Science and Technology Commission [CSTC2018jcyjAX0246].

References

1. A. Downes, C. Thornhill, R. H. Mallard, Commun Dis Public Health 1, 206-207 (1998).
2. Y. W. Tang, N. M. Ellis, M. K. Hopkins, D. H. Smith, D. E. Dodge, D. H. Persing, J Clin Microbiol 36, 3674-3679 (1998).
3. A. Innings, M. Krabbe, M. Ullberg, B. Herrmann, Journal of Clinical Microbiology 43, 5983-5991 (2005).
4. E. Garza-Gonzalez, G. M. Gonzalez, A. Renteria, W. Cruz-Pulido, G. Rivera, V. Bocanegra-Garcia, Clin Microbiol Infect 16, 607-612 (2010).
5. M. Vos, C. Quince, A. S. Pijl, M. de Hollander, G. A. Kowalchuk, Plos One 7, (2012).
6. I. B. Kahla, M. Henry, J. Boukadida, M. Drancourt, BMC Res Notes 4, 423 (2011).
7. D. Pu, R. Pan, W. Liu, P. Xiao, Electrophoresis 38, 876-885 (2017).
8. L. Varadi, J. L. Luo, D. E. Hibbs, J. D. Perry, R. J. Anderson, S. Orenga, P. W. Groundwater, Chem Soc Rev 46, 4818-4832 (2017).
9. D. Pu, P. Xiao, RSC Advances 7, 40141-40151 (2017).
10. D. Pu, Y. Qi, L. Cui, P. Xiao, Z. Lu, Anal Chim Acta 852, 274-283 (2014).
11. T. Matsumoto, M. Sugano, Rinsho Byori 61, 1107-1115 (2013).
12. R. J. Clifford, M. Milillo, J. Prestwood, R. Quintero, D. V. Zurawski, Y. I. Kwak, P. E. Waterman, E. P. Lesho, P. Mc Gann, PLoS One 7, e48558 (2012).
13. J. Isaksson, M. Rasmussen, B. Nilson, L. S. Stadler, S. Kurland, L. Olaison, E. Ek, B. Herrmann, Diagn Micr Infec Dis 81, 240-245 (2015).
14. A. B. Teixeira, J. Barin, D. M. Hermes, A. L. Barth, A. F. Martins, J Clin Lab Anal 31, (2017).
15. P. Priyadarshini, K. Tiwari, A. Das, D. Kumar, M. N. Mishra, P. Desikan, G. Nath, Int J Tuberc Lung Dis 21, 214-217 (2017).
16. E. S. Haas, J. W. Brown, Nucleic Acids Res 26, 4093-4099 (1998).