Prediction of PCR amplification from primer and template sequences using recurrent neural network

Kotetsu Kayama1, Miyuki Kanno1, Naoto Chisaki1, Misaki Tanaka1, Reika Yao1, Kiwamu Hanazono1, Gerry Amor Camer1,2 & Daiji Endoh1,2*

We have developed a novel method to predict the success of PCR amplification for a specific primer set and DNA template based on the relationship between the primer sequence and the template. To perform the prediction using a recurrent neural network, the usual double-stranded formation between the primer and template nucleotide sequences was herein expressed as a five-lettered word. The set of words (pseudo-sentences) was placed to indicate the success or failure of PCR targeted to learn recurrent neural network (RNN). After learning pseudo-sentences, RNN predicted PCR results from pseudo-sentences which were created by primer and template sequences with 70% accuracy. These results suggest that PCR results could be predicted using learned RNN and the trained RNN could be used as a replacement for preliminary PCR experimentation. This is the first report which utilized the application of neural network for primer design and prediction of PCR results.

PCR primers have been traditionally designed by thermodynamic interaction with the desired templates1,2. Primers are designed to increase two respectively significant base sequence specificity and reasonable GC content indicators. The high specificity can prevent mispriming in regions other than the target region, and the GC content of a primer is a major factor in determining the annealing temperature (Tm). Maintaining the Tm value optimally affects the amplification efficiency of primers being used3. The reference scaffold for primers with high PCR success has been determined based on the result of trials of up to around 1990 thermodynamic calculations1. Indices such as discontinuity of the same base are also empirically determined. The Tm value, which explains the specificity of binding to the template and possibly with the primer dimer, among others, are evaluated to determine the appropriate primer pair on each template. With this, some of the proposed software has been designed. The most frequently used primer design software include PrimerSelect4, Primer Express (Applied Biosystems Primer Express Software Version 3.0 Getting Started Guide, 2004), Primer Premier (http://www.premierbiosoft.com/primerdesign/index.html), OLIGO 75, Primer36, and OMP7. Of these primer design softwares, Primer3 software provides both a primer design on the Web and a local program (Primer3_core) that becomes a standard for PCR primer design. In particular, Primer3 added some thermodynamic findings in 2007 and 20122,7. Its revision in 2012 provided an added knowledge about DNA duplex stability8 which incorporated an algorithm for designing primers on the target9. This enabled a primer design even in the boundary regions of exons.

Current primer design techniques allow the design of primers that amplify the subject sequence with high probability resultant of combining thermodynamic theory alongside the experience of many researchers. However, it has not been designed to make predictions about a nucleotide sequence that is said to be "not amplifying" a known template. In some previous cases where amplification was performed with an unexpected template in a PCR experiment, knowledge-feedback was unfortunately not documented. Earlier contributions on PCR primer designs have incorporated these into modifications of thermodynamic laws before being compacted to a primer design software3,5. To indicate the presence of a particular DNA or RNA sequence, there is always a need to predict that no PCR will occur at sequences that are not of interest or prime importance. In pathogen detection, PCR primers are selected based on several preliminary experiments to confirm that PCR can predictably occur only with a specifically targeted pathogen10–13. Since false positives pose a major problem in detecting many

1Department of Radiation Biology, School of Veterinary Medicine, Rakuno Gakuen University, 582 Midori-machi Bunkyo-dai, Ebetsu 069-8501, Japan. 2University of Eastern Philippines, College of Veterinary Medicine, 6400 Catamaran, Northern Samar, Philippines. *email: dendoh@rakuno.ac.jp
pathogens including COVID-19, it is important to develop a method for predicting false positives (https://www.biotechniques.com/covid-19/false-negatives-how-accurate-are-pcr-tests-for-covid-19/). Hence, if specificity of a primer pair can be predicted from nucleotide sequences of primers and templates, hindrances including false-positives can be readily corrected resulting to an accelerated research process.

To enable PCR results from the base sequences of primers and templates, it is necessary to comprehensively evaluate various relationships between primers and a template. We focused on machine learning as a technique for predicting a PCR result from various primer-template relationships. Machine learning has been achieving positive remarkable results particularly on data analysis. In machine learning, the results of input data can be predicted based on various factors without assuming a unified theory, through optimizing the coefficients of the perceptron network. When a base sequence of a PCR primer and a template is set as a target of machine learning, it is expected that a character string anchored on the base sequence may be suitable for a target of learning. One of the methods for machine learning that has been successful in processing languages used in human research is a recurrent neural network (RNN). As a feature of natural language processing, sentences can be classified based on the frequency and arrangement of words. It has been considered that if the relationship between the primer and the template related to PCR could be replaced with a word, then, it could be a target of natural language processing.

When predicting PCR results, it is necessary to generate learning data for machine learning from the base sequences of the primer pair and the template. The PCR results obtained in the experiment can be used as the correct answer in supervised learning. Factors influencing PCR by primer pairs and templates consist of different relationships and their positions such as dimers, hairpin loops, and partial complementarity. To comprehensively evaluate atypical relationships of these different factors, it was inferred that the Recurrent Neural Network (RNN), which predicts the meaning of sentences from the frequency and arrangement of words, is optimal. With this, we aimed to build a supervised learning method. We generated a pseudo-sentence from a relationship of a primer pair and a template. The PCR results could be learned using RNN that is a supervised learning method of a natural language. When this learning method is used, it is expected that new supervised learning can be performed even when the results differ due to variable settings of annealing temperature, among others, for the combination of a specific primer pair and a template. In this study, we report the prediction of PCR results by supervised learning.

Materials and methods

Assumption. To create training data on RNN, the entire PCR reaction was schematically planned. (Fig. 1). Primer binding to the template is not limited to its full length and is assumed that only a part of 3’ may bind (Fig. 1B). Hairpin structure of the primer and its dimer are assumed to be formed before binding the primer to the template (Fig. 1B). Thus, it is assumed that DNA synthesis occurs from some hairpin structures and dimers. As DNA synthesis from partially bound primers proceeds, PCR products that are completely complementary to the primers began to be synthesized (Fig. 1C). Eventually, most PCR products become completely complementary to the primers (Fig. 1D).

To express the relationships of these schemas as words, we decided to express the hairpin, primer dimer, primer-template bond, and primer-PCR product bond as words. The strength of the primer-template bond on the forward and reverse flanks greatly influences the establishment of the PCR reaction. For combinations that
are not of the original primer-template, the binding position needs to be determined by PCR from the possible binding of multiple primer-templates. With this, we constructed the words for the learning RNN.

**Templates for PCR.** A part of the 16S rRNA nucleotide sequence (v6-v8) (Supplement 1 Table 1) was synthesized by OE-PCR (Supplement 1 Table 2) for 30 phyla as templates for PCR model experiments. Of the 30 phyla, 16S rRNA sequences in Firmicutes were synthesized into two genera, the Bacillus and the Calditerricola. These sequences were significantly different in v6-v8. Thirty-one double-stranded DNAs with 435 to 481 bases were prepared as a template for PCR model experimentation utilized the standard thermodynamic index.

**PCR experiment as basic data for primer design using RNN.** Design of primer sets for preliminary learning of RNN. We designed 72 sets of PCR primers capable of amplifying 31 DNA templates, according to the specifics of the primer sequence to the specific template and the amplification size of about 100–150 (Table 1). In a preliminary trial when primers were designed using Primer3 primer-design software, all primers amplified all 31 templates (data not shown). From its result, we designed 72 sets of PCR primers at this stage ignoring some of the conventionally known annealing temperatures and some indicators such as avoiding single base repetition. The size of the primers was set to 19–22 bases. The most important index is high homology to the target template and low homology to others.

We also designed 54 sets of phylum-specific primers, which were designed based on analysis with preliminary test primers (Table 2). As a design method, a plurality of primer candidates was firstly extracted from the template sequence, and a combination of the extracted primer candidates was used as a primer pair candidate. A primer pair for which PCR is expected to occur only in a specific bacterial phylum by RNN was determined as a primer set for a test experiment.

**PCR amplification experiments.** Using the 72-primer sets for learning and validation of RNN and 54-primer sets for testing RNN, we tried to amplify all 31 templates. PCR was carried out using 2× GoTaq Green Hot Master Mix (Promega) for a total of 3,906 PCRs with 31 templates and 126 (72 plus 54) sets of primers. The PCR solution contained 0.5 µM primer, 100,000 copies of the template, and was adjusted to 1× GoTaq Green Hot Master Mix by adding water and 2× GoTaq Green Hot Master Mix. After adjustment, the PCR solution was subjected to denaturation at 95 °C for 2 min and followed by 33 cycles at 95 °C for 30 s, 56 °C for 30 s, 72 °C for 30 s, and followed by incubation at 72 °C for 2 min. After cooling to 8 °C, it was stored at 4 °C until processed in agarose gel electrophoresis. The PCR products were electrophoresed using 1.5% agarose in 1× TBE buffer at 100 V for 40 min. The agarose gel was stained in 1 µg/ml Ethidium Bromide solution and photographed under UV.

**Symbols for RNN learning.** The data for RNN learning consisted from a symbol (Table3) generated from the hairpin structure of the primer, the primer dimer, and the homology between the primer and the template, and multiple 5-character codes (pentacode) generated from the symbol (Fig. 2). The correct answer data for RNN was the PCR result for each primer set and template. Since the RNN is optimized for learning natural language sentences, which were composed of words, the generated pentacode is called a pseudo-word, and the pentacode listed according to the nucleotide sequence of the template is called a pseudo-sentence. Specific design methods are described in the creating pseudo-words and pseudo-sentences section.

**Creation of pseudo-words and pseudo-sentences from the relationship between primers and templates.** For hairpins and dimers, DNA synthesis from the complementary region was predicted and the synthesized primers were added to the primer set. For the complementary region between the hairpin, dimer, primer-template, and primer-PCR product, characters corresponding to the complementary base pair were set according to the pentacode. As a general rule, the homology between the primer sequence divided into 5 to 22 bases and the template sequence was confirmed, and when the number of bases in Supplement 2 Table 1 was the same (about 80%), a pseudo-code was generated (Fig. 2D). Regarding the homology, area to be generated as a pseudo-code, the pseudo-code was determined by referring to Table 3 for the entire homology, and all lower-case pseudo-codes were generated (Fig. 2E).

Many primer set-template combinations have multiple complementary regions that require priming positions to be determined. Since the complementary region for which such a priming position needs to be determined is short enough, the most stable combination of complementary regions is expected to be the priming position. To determine the most stable complementary region, the combination of complementary regions with the minimum Gibbs energy was set as the priming position (Fig. 2F). The Gibbs energy was calculated according to the formula of DG = DH−TDS by sequentially calculating the entropy and enthalpy of the two bases of the primer and the two
| Primer no. | Primer name | Sequence |
|-----------|-------------|----------|
| 1         | aim_1f      | GTCCAGGGCTTCTACACATGCTA |
|           | aim_22r     | TGTTACAACTTCTCTAGACGTG |
| 2         | aim_71f     | AGCGCAACCCCTACACCTTGTGT |
|           | aim_94r     | GGGACGGATTTGAGATTAG |
| 3         | aim_122f    | TTCAGATTGGCACATCGTAAGGA |
|           | aim_94r     | GGGGGGATTTTGAGATTAG |
| 4         | aim_94f     | CTAATCTCAAAATCCGGTCCC |
|           | aim_143r    | CTTCCAGGATTTTACCCATTAGT |
|           | aim_94f     | CTAATCTCAAAATCCGGTCCC |
|           | aim_147r    | CTTACCAACCTCTAGAGATTTC |
| 5         | aim_161f    | GAGGTTGGACGGAATCCGAGAAAA |
|           | aim_194r    | CTTACCAACGATACCTTGAACGA |
| 6         | aim_263f    | CAAATCCAGAAAGGCGCCCTC |
|           | aim_250r    | ACCAGCCTTCCGCGGCTCC |
|           | aim_348f    | GTTGGCCCTAGCAATTAGATCT |
|           | aim_282r    | TGCTGCCCTCTGTCTCAAGCAT |
| 7         | aim_386f    | GCTGAGGACTCTTAATTGAACGT |
|           | aim_394r    | AGACAGCTTTAAAGGATTTCCT |
|           | aim_386f    | GCTGAGGACTCTTAATTGAACGT |
|           | aim_461r    | CGGATCGGAGACTGAGACAGCTT |
| 8         | aim_394f    | GGAATCCCTTAAAGAGCTTCT |
|           | aim_436r    | CGAGGGCTCTTGGGTACCTCCCT |
|           | aim_468f    | GGGGGGACCTAACCAAAACTC |
|           | aim_515r    | CTTCAAGAATCTCGGGTGCCAGAC |
| 9         | aim_555f    | AGCGGACTCGCCGGCAAAGGG |
|           | aim_562r    | GGGCGCCACTTCTTTGCGATTAG |
|           | aim_599f    | GTGACTACAGGAGTAAGCAGAAAC |
|           | aim_652r    | CCGCCGAGGGCGAGTTGGGTCAC |
| 10        | aim_764f    | CTTATCCCTTGTGGCACGGCGT |
|           | aim_744r    | CGACAGCATTTATGACGGTCGCC |
| 11        | aim_812f    | CTGGCAGTAGATAAACTGGAGGA |
|           | aim_744r    | CGACAGCATTTATGACGGTCGCC |
| 12        | aim_842f    | TCTCATAAACCCGGTTCTCAGTT |
|           | aim_848r    | TGTTAAACACTCTCTGGTGAGTG |
| 13        | aim_1016f   | CTAATCGGAAAAAGCCGGGCCCTC |
|           | aim_1036r   | ATGAAATACACCTTGCGGCGCT |
|           | aim_1209f   | GTGCTCGGTAGTACACCTGCTT |
|           | aim_1159r   | ATTGCTCGGGCACATGTAGCAGT |
| 14        | aim_1248f   | CGCGTGAGCCGCGGAGGAAGG |
|           | aim_1193r   | CGCGCCTAGTGCTGATTACGGG |
| 15        | aim_1175f   | TCGCCTAAACCTGGTCTCAGTG |
|           | aim_1264r   | CCGCGCGGAGTGCTTCGGCCCT |
| 16        | aim_1177f   | GCCCTAAACCTGGTCTCAGTG |
|           | aim_1264r   | CCGCGCGGAGTGCTTCGGCCCT |
|           | aim_1316f   | CTAGTGGGACAGCAGGAGTAAT |
|           | aim_1285r   | TGCAATCCGAACTAAGACAAAGG |
| 17        | aim_1332f   | CGGAGTAACCCCGGAGGAAGGT |
|           | aim_1276r   | AGGTGTTGTGAGGTGGCTCAGT |
| 18        | aim_1285f   | CTTGCTTCTAGTTCGGAATTGCA |
|           | aim_1293r   | GCTCCTGCGCAAACCAGACTTTC |
| 19        | aim_1316f   | CTAGTGGGACAGCAGGAGTAAT |
|           | aim_1285r   | TGCAATCCGAACTAAGACAAAGG |
| 20        | aim_1332f   | CGGAGTAACCCCGGAGGAAGGT |
|           | aim_1276r   | AGGTGTTGTGAGGTGGCTCAGT |
| 21        | aim_1285f   | CTTGCTTCTAGTTCGGAATTGCA |
|           | aim_1293r   | GCTCCTGCGCAAACCAGACTTTC |
|           | aim_1401f   | TGAGGGTCGCTTAAGTGCCAA |
|           | aim_1368r   | GCTAAGCTGCTTCGTAACCCC |
|           | aim_1401f   | TGAGGGTCGCTTAAGTGCCAA |
|           | aim_1383r   | TTGGGATTAGCATACGGTCAC |
| 22        | aim_1413f   | GAGAGGTCGCTTAAGTGCCAA |
|           | aim_1368r   | GCTAAGCTGCTTCGTAACCCC |
|           | aim_1413f   | GAGAGGTCGCTTAAGTGCCAA |
|           | aim_1368r   | GCTAAGCTGCTTCGTAACCCC |

Continued
| Primer no. | Primer name | Sequence                  |
|-----------|-------------|---------------------------|
| 29        | aim_1447f   | AGGTCATGCTGAGGACTCTGGA    |
|           | aim_1368r   | GCTAGCGCTCTGCTGACCCCC    |
| 30        | aim_1447f   | AGGTCATGCTGAGGACTCTGGA    |
|           | aim_1391r   | TTTGATCCGAACTGAGAGAGGA    |
| 31        | aim_1391f   | TCCCTCTCTAGTTGGATGCAAA    |
|           | aim_1465r   | CCAAAGAACAGACTCTGAGGTT    |
| 32        | aim_1549f   | GGTGATGGCGGGTTACACAG      |
|           | aim_1504r   | GATTGGTCTCAGGCTGAGG       |
| 33        | aim_1504f   | CTGCTACAGTGCCGGGAAATG     |
|           | aim_1550r   | TGGTCTGCCGGACTTCCGAGG     |
| 34        | aim_1758f   | GCCAATACACAGATGTTGAAAAT   |
|           | aim_1775r   | TACCGCTCTGTAAGTTGAGG      |
| 35        | aim_1775f   | TCGTAAAGTGCTGCTGAGG       |
|           | aim_1775r   | TACCGCTCTGTAAGTTGAGG      |
| 36        | aim_1770f   | CTGTTAAGTGCTGCTGAGG       |
|           | aim_1811r   | CTACCCTAGAATGCGCTTCTCT    |
| 37        | aim_1948f   | CAAAGGGCGAGGGATAGATGAT    |
|           | aim_1984r   | ATGAGCGGTAGCTGATGCCCATT   |
| 38        | aim_2085f   | AGTACAGAGGTTAAGGCAATCG    |
|           | aim_2138r   | AACGTATTACGGCGGGTATTAGGC  |
| 39        | aim_2109f   | GATGGACGCAAATCTCTAAAGCT   |
|           | aim_2164r   | TCAACCGACTTAAAGTAAACCA    |
| 40        | aim_2292f   | GGTAAAGTCCCCCTAAGGACCGA   |
|           | aim_2247r   | ATGACCTGGAGCCTAGCAACCG     |
| 41        | aim_2547f   | TCGGTAGACTGAAAGTGGATAC    |
|           | aim_2581r   | TACCGTTAGGCGCCCTGCTACTCA  |
| 42        | at2_1242_f  | TACCTTGTCTAACGAGACCTGCC   |
|           | at2_1242_r  | CGAATCGAGACCAACTTTTACAG   |
| 43        | at2_100_f   | AGCAGGCGCGAGGAAGGAGG      |
|           | at2_100_r   | ACCCGCGGAAACTATTCACC      |
| 44        | at2_1213_f  | CCTAAACCTCTGCTGTGGTGCAG   |
|           | at2_1213_r  | TAGCTCCGGGAGCTTCCGATGAA   |
| 45        | at2_1325_f  | TAAAGGGGACTGCCCGGATAAC    |
|           | at2_1325_r  | GCGCTTTCTGAGATCTGCTAC     |
| 46        | at2_6_f     | CAAGTGCAGCGGAGAGATTT     |
|           | at2_6_r     | GGTATTACCCATCTCTTCGGAT    |
| 47        | at2_1194_f  | GGCGGTGACGCTATGCTAAATCC   |
|           | at2_1194_r  | TTGCGGTGATCCTACTTGGAGT    |
| 48        | at2_1315_f  | CGTTGTCTAGGGTGCAAAAGCTCF  |
|           | at2_1315_r  | GGCGCTGCGGCGGACTTCCGAGT   |
| 49        | at2_1147_f  | CGCCGTGACCGCCCCCGGAGAGG   |
|           | at2_1147_r  | CACTGTACCGACAGTTACGCGA    |
| 50        | at2_23_f    | GAGACTGCCTGCGGACAAACC     |
|           | at2_23_r    | AGTTGAAGACTCCCAATCCGGA    |
| 51        | at2_1244_f  | GCAAATACAAACAAGTGGCAAAT   |
|           | at2_1244_r  | TACCACTCTCATAGTTGGACG     |
| 52        | at2_1125_f  | ACCGCTGCAACCCCCCGGAGGTT   |
|           | at2_1125_r  | TTGGCCGGCGTCTCCTTGGGAGT   |
| 53        | at2_1238_f  | GGCCACAGCTTGTTGACCGGCGT   |
|           | at2_1238_r  | GGCCATAAGGGGGCGACGATACCT  |
| 54        | at2_1166_f  | CCGGAGTTAATCCCGGAGGAAGGT  |
|           | at2_1166_r  | TGCAATCCCGAATAGACCAAGG    |
| 55        | at2_1143_f  | TCCCGCCCGTACCGCGGAGGGA    |
|           | at2_1143_r  | CACTGTAGACCGTATTAGGCGA    |
| 56        | at2_1124_f  | ACCGCTGCAACCCCCCGGAGGTT   |
|           | at2_1124_r  | AGCGCACCAGCTTCTAGTGCAA    |

Continued
bases of the template at the complementary position, assuming an annealing temperature of 56 °C. Therefore, after calculating for all combinations of two complementary bases, the total value was minimized, and the complementary positions of forward and reverse, which are separated by 100 bases or more, were set as the priming positions. Using reference numerical values, complementary dimer set calculations for entropy and enthalpy were done where their original and our extrapolated values were used (Supplement 2 Table 4). The pseudo-code for the complementary position, which was predicted to be the priming position, was converted to uppercase (Fig. 2G). Homologous positions of 6 bases or more were searched for hairpins and dimers, and pseudocodes were generated for the corresponding homologous regions (Fig. 2H). For the pseudo-code sequence generated between the primer and the template, 5 characters were sequentially extracted from the 3′ end of the primer to obtain a pentacode. The pseudo-code was generated and placed from a set of primers and a template was used as pseudo-sentences of the primer set-template (Fig. 2K). Pseudo-sentences were generated for all primer and template combinations and used as learning data during machine learning.

| Primer no. | Primer name | Sequence |
|------------|-------------|----------|
| 57         | ai2_1284_f  | ACGAGACTGCCGCTGGGTTAACCA |
| 57         | ai2_1284_r  | AGCTTTAAGATTGTGCACTCAC |
| 58         | ai2_1286_f  | CGTGCTGGGTAAAAAGCGAGAGG |
| 58         | ai2_1286_r  | GAACCTGCGGGCAGCTTTTAAAG |
| 59         | ai2_1090_f  | AAGGGAGACGTCGAGTCGATAAA |
| 59         | ai2_1090_r  | TCCATTCGCACTACGACATAAC |
| 60         | ai2_1142_f  | GTGTCGCTGTAGTAAAGGGTTG |
| 60         | ai2_1142_r  | CAACCTGAAGCTCCAGGGGGGG |
| 61         | ai2_1195_f  | GCGGCGTACGGTATGCTAAATCC |
| 61         | ai2_1195_r  | CCGTACGAGTACGCTGAGG |
| 62         | ai2_101_f   | GTGCTGCTGCACTGGTCC |
| 62         | ai2_101_r   | CTCCTTCCTCCGCGCTT |
| 63         | ai2_1189_f  | GCTGCTAGATGTGAGGAAGGT |
| 63         | ai2_1189_r  | TTGCACTACGACTGAGAGGA |
| 64         | ai2_1088_f  | TATATCTTCAGGGCCCAGG |
| 64         | ai2_1088_r  | TCCATTCGCACTACGACATAAC |
| 65         | ai2_1192_f  | GGGGTTACAGAAAGCGACTAGC |
| 65         | ai2_1192_r  | CTGGCCGCTACGTATCCAGG |
| 66         | ai2_1303_f  | GCCATAACGGCGTGAATACGTT |
| 66         | ai2_1303_r  | TCTCATGGTGATGCTGAC |
| 67         | ai2_1102_f  | AAGTTGGGACAGTCTAAGGTGAC |
| 67         | ai2_1102_r  | TCTGCACGCTCTTTTCTAGCAG |
| 68         | ai2_54_f    | CGGGTGGATACAGCGTATCCTAA |
| 68         | ai2_54_r    | TCTGATTCGGCGCTACGTATCAT |
| 69         | ai2_1275_f  | TGATATGGGAGCGAATCCCCAAA |
| 69         | ai2_1275_r  | GTCTGCTCCCTGCCAAGAGGT |
| 70         | ai2_1327_f  | CTGAGCGGAATCTGAGAAAAGC |
| 70         | ai2_1327_r  | TGTGCCGCTGTTGGGACCACGGC |
| 71         | ai2_10_f    | CTGGCCGCGTGTGGATAAGACA |
| 71         | ai2_10_r    | ATGGGCATATCCCCACCTCAG |
| 72         | ai2_1071_f  | AGCGATGCGCCACCGCGCAAGGG |
| 72         | ai2_1071_r  | CCTGCAGCAGGGCTCCCGGCGA |

Table 1. Primer sets for the main experiment. Primer pair number, primer name and base sequence (5′→3′) used in the experiments for RNN-training are shown. Primers with the same primer pair number are used as a set of primers.
| Primer set no. | Primer name | Primer sequence |
|---------------|-------------|-----------------|
| 1             | 1_f_180     | AACGCGCTCGGAGGCCGCTGTA |
|               | 1_r_369     | CCCACAAGGGTATGGCCCATC |
| 2             | 1_f_180     | AACGCGCTCGGAGGCCGCTGTA |
|               | 1_r_397     | ATCGGCCATCCACATCTACGTA |
| 3             | 1_f_180     | AACGCGCTCGGAGGCCGCTGTA |
|               | 1_r_400     | CCAATGCCGATCCCCACCTTG |
| 4             | 1_f_180     | AACGCGCTCGGAGGCCGCTGTA |
|               | 1_r_408     | ACTTCTGTCACATCGCCGAT |
| 5             | 1_f_180     | CTGCCAGCCTGTGAGGGGTA |
|               | 1_r_369     | CCCACAAGGGTATGGCCCATC |
| 6             | 1_f_180     | CTGCCAGCCTGTGAGGGGTA |
|               | 1_r_397     | ATCGGCCATCCACATCTACGTA |
| 7             | 1_f_180     | CTGCCAGCCTGTGAGGGGTA |
|               | 1_r_400     | CCAATGCCGATCCCCACCTTG |
| 8             | 1_f_180     | CTGCCAGCCTGTGAGGGGTA |
|               | 1_r_408     | ACTTCTGTCACATCGCCGAT |
| 9             | 2_f_50      | TCAGTTGCCGACTGTAGG |
|               | 2_r_342     | TGGCAAAAGACCTGCGGG |
| 10            | 4_f_251     | CTAAGGCCACCCGATCACA |
|               | 4_r_395     | CTCTTGCCGACTGCGGG |
| 11            | 4_f_251     | CTAAGGCCACCCGATCACA |
|               | 4_r_403     | TGGCGAGGCTCTTGCGGCTGTA |
| 12            | 5_f_0       | TGCTGCGGAGCCCATACGCA |
|               | 5_r_228     | CCCCTAGGGTCTGCTCC |
| 13            | 5_f_233     | GAGGGGAGGCGAGAACCCGTA |
|               | 5_r_410     | TGGGAGGCTTTGGATAGCGA |
| 14            | 6_f_66      | GCTTCAGCGAATATCTCCTC |
|               | 6_r_211     | GGCGCATGGTTAAGGATG |
| 15            | 6_f_211     | CCAATCCCTAAACTATTGCC |
|               | 6_r_363     | AGACGCACCTGAGCCTCCTG |
| 16            | 6_f_211     | CCAATCCCTAAACTATTGCC |
|               | 6_r_385     | TACTAATCACACCTAGGCC |
| 17            | 7_f_205     | AAATCCCTAAAAGGCTGTCC |
|               | 7_r_347     | AGCGCTTCTTGGTACTCCTG |
| 18            | 8_f_78      | ACTGCCAGATCAACTGGGA |
|               | 8_r_349     | TGGCTTCAGATCTTGCGGGT |
| 19            | 8_f_78      | ACTGCCAGATCAACTGGGA |
|               | 8_r_359     | TCCCTGCGGTTGCGCTCAGA |
| 20            | 9_f_87      | GACCTGCCGCGAAAGCGGG |
|               | 9_r_181     | GTGCGCGGTTGCGCTCAGA |
| 21            | 10_f_226    | TCCCTAAAAGCATCCCTAG |
|               | 10_r_381    | AGCGGAAGTGGCGCTCAGA |
| 22            | 12_f_190    | GCCGATACAAAGGAGACCG |
|               | 12_r_395    | TAAGCGGCCCTCCGGAAGGT |
| 23            | 12_f_216    | CGAGAGCAAGGCGACCTCAT |
|               | 12_r_395    | TAAGCGGCCCTCCGGAAGGT |
| 24            | 12_f_236    | AAAATGATGCTGATCTCCGGA |
|               | 12_r_395    | TAAGCGGCCCTCCGGAAGGT |
| 25            | 13_f_203    | GGTGAAAGGACTGCAAGAC |
|               | 13_r_397    | CTCCAAAAAGGGTTACCAC |
| 26            | 14_f_64     | GCAGGAGGAGGCGCTCCCGAGA |
|               | 14_r_341    | TAGAGCACTCCCCTCCTC |
| 27            | 15_f_60     | TGCCGAAACCCTCGGATA |
|               | 15_r_349    | TTCCTCTGGGTTAGGACAC |
| 28            | 18_f_17     | TGCTCGGATGATGAGGCTG |
|               | 18_r_177    | GATCTGCAGCTGAGACACG |
| Primer set no. | Primer name | Primer sequence |
|---------------|-------------|-----------------|
| 29            | 18_f_173    | CTAACGCTGCTTCAGTGCA |
|               | 18_r_330    | TCCCGACTGGGTATGAC |
| 30            | 19_f_83     | TGGTGGGCAGCGGAGTAA |
|               | 19_r_177    | AGTGCGCTCCGGTTGCTT |
| 31            | 19_r_90     | ACACCCGAAGTAGTCCGAG |
|               | 19_r_371    | CTATCCGAAGATTGGGTCAC |
| 32            | 19_r_196    | TCGCGAGTGAGCGCAACCT |
|               | 19_r_346    | TCTGGAAACCCGACTTTCG |
| 33            | 19_r_196    | TCGCGAGTGAGCGCAACCT |
|               | 19_r_371    | CTATCCGAAGATTGGGTCAC |
| 34            | 20_f_225    | ATCCCAAAAATCTCTCTCAG |
|               | 20_r_382    | GAGCAGCTTGGGATACGT |
| 35            | 21_f_61     | GGTTAAGCGGGGTACCTC |
|               | 21_r_211    | ACCACGACGGTATGGGG |
| 36            | 21_f_61     | GGTTAAGCGGGGTACCTC |
|               | 21_r_217    | ATCTGCACCAACAGGGTT |
| 37            | 21_f_61     | GGTTAAGCGGGGTACCTC |
|               | 21_r_227    | GCAACCCCTCATACTGCAA |
| 38            | 21_f_179    | AATGGGCTGCAACCCGGA |
|               | 21_r_359    | GTAGCTGGGGGACTCCGA |
| 39            | 21_f_216    | AAACCCCTGTGCTGGGTGCA |
|               | 21_r_348    | GACTCCGATGAAACCGACT |
| 40            | 23_f_151    | ATGACGCTAGTACCTGTCG |
|               | 23_r_437    | CCCCCCTCCACAGTGTCC |
| 41            | 24_f_84     | CTCGCAAGTGAATGTGAGG |
|               | 24_r_360    | CTTCGCGTATGCAACAGGGT |
| 42            | 26_f_74     | AGACTCCCGTGTTAAGGGG |
|               | 26_r_169    | TCACATATCGGCTGGCCTT |
| 43            | 26_f_74     | AGACTCCCGTGTTAAGGGG |
|               | 26_r_358    | CTGCAACAGGCTTGGGGCA |
| 44            | 28_f_165    | AATGGGCGCGACAGAGGCTT |
|               | 28_r_333    | CCCCCGCTTGGTGGCTGTA |
| 45            | 28_f_165    | AATGGGCGCGACAGAGGCTT |
|               | 28_r_397    | ACTTAGTGCCCCATACCGG |
| 46            | 28_f_170    | GGCGGAGAGACAGTTGGCTAG |
|               | 28_r_339    | GGATGCCGCCGCTTGTTGGTG |
| 47            | 28_f_173    | GGACAGAGGCTTGACTAGGCT |
|               | 28_r_333    | CCCCCGCTTGGTGGCTGTA |
| 48            | 28_f_192    | TGCAAGTCTGATCTAATCGC |
|               | 28_r_333    | CCCCCGCTTGGTGGCTGTA |
| 49            | 28_f_192    | TGCAAGTCTGATCTAATCGC |
|               | 28_r_397    | ACTTAGTGCCCCATACCGG |
| 50            | 28_f_204    | CTAATCGGAAAACCCGTCC |
|               | 28_r_397    | ACTTAGTGCCCCATACCGG |
| 51            | 29_f_237    | GGGAATCTCGAAGAGCGCCTC |
|               | 29_r_437    | CCCAGTGCGACGCCATACCA |
| 52            | 30_f_128    | CAAATGCTACGGAGAAGGGG |
|               | 30_r_306    | TCCGCGAGTGCGCCAAATTCC |
| 53            | 30_f_128    | CAAATGCTACGGAGAAGGGG |
|               | 30_r_328    | CCACAAGGTGGTGGATACAG |
| 54            | 30_f_128    | CAAATGCTACGGAGAAGGGG |
|               | 30_r_342    | TCCGCGAGTCTCTCCGAAA |

Table 2. Primer sets for the test. Primer pair number, primer name and base sequence (5′- > 3′) used in the experiments for RNN-test are shown. Primers with the same primer pair number are used as a set of primers.
| Base pair (primer base–template base) | Primer-template characters | Initial stage | Middle stage | Primer hairpin or dimer |
|-------------------------------------|---------------------------|---------------|--------------|-------------------------|
| A-T, T-A                            | a                         | f             | p            | u                       |
| C-G, G-C                            | b                         | g             | q            | v                       |
| A-A, A-G, G-A, G-G, C-C             | c                         | h             | r            | w                       |
| T-T, C, C-T                         | d                         | i             | s            | x                       |
| C-A, C-G, G-T, T-G                  | e                         | y             | t            | y                       |

Table 3. Base pair characters for sense or antisense direction. A symbol for generating pseudo-words for RNN learning. The codes are set in the nucleotide duplex on each base pair at the complementary position. Mismatches at base pairs such as A-A, T-T and C-A may appear within the partially complementary region. Base-pairs are grouped based on influence for stability of partially complementary strands.

Scripts for pseudo-sentence generator. A Ruby and Python script were used to generate pseudo-sentences in the order shown in Fig. 2 (Supplement 3, List 1–9). The Ruby script read the structure of the template base sequence, primer base sequence, and primer set, and generated pseudo-sentences according to the order shown in Fig. 2. SeqKit ([https://bioinf.shenwei.me/seqkit/](https://bioinf.shenwei.me/seqkit/), v0.14.0) was used to search for homology between the primer and the template. The pseudo-sentences generated for each template-primer set were first categorized by PCR results, and each was categorized into 5 groups. One of the five groups was not used for learning as a group to verify RNN learning but was used to predict the prediction accuracy for each epoch.

We noted that a particular primer set produced many positive PCR results and organized the group to disperse its effects. Five groups were randomly constructed for each PCR positive and negative results after collecting the results for each template. To divide the overall result into 5 groups, the primer-template pair data, which is the unit of data, was divided so that the total number was even for each group. When we equalize the ratio of PCR positives and negatives, the acquired data is adjusted so that the numbers are even at the stage of collecting the results for each template (undersampling).

Axlsx ([https://github.com/randym/axlsx](https://github.com/randym/axlsx), v3.0.0) was used for colorizing spread sheets (Tables 4, 7). MatPlotLib ([https://matplotlib.org](https://matplotlib.org), v3.3.3) was used for creating line-graphs on epochs-accuracy (Fig. 4). GnuPlot ([http://www.gnuplot.info](http://www.gnuplot.info), v5.4) was used to create the scatter plot for Gibbs energies (Fig. 5).

Learning results. The PCR results performed with the annealing temperatures set at 56 °C were set as pseudo-texts generated from each primer-template set and were trained by RNN. For its learning, the pseudo-sentence created for the combination of primer and template was used as input data, and the PCR results were arranged as a teacher. For the RNN, an RNN-Long short-term memory (LSTM) module of PyTorch ([https://pytorch.org](https://pytorch.org), v1.7.1) was used. Python scripts for learning pseudo-sentences and extracting prediction results were written based on the scripts published in a book (Shinqiao Du, "Can be used in the field! Introduction to PyTorch development Creation of deep learning model and implementation in application", Shosuisha; 2018/9/18 in Japanese). After reading the pseudo-sentences and PCR results of each primer-template set, RNN generated a decision algorithm that matched the output results for all input pseudo-sentences (learned algorithm) (Fig. 3). As the negative control of sentences, randomly selected nucleotide pentamers were aligned as nonsense pseudo-sentences.

The prediction accuracy of the generated trained algorithm was confirmed by split verification (cross validation). The primer pair-template sets were divided into five groups, and the RNN was learned using four groups among them and the learning. The remaining one group was not used as learning data but was utilized as verification data. Verification was made during the learning steps.

When evaluating the prediction by RNN, whether the expected PCR band was found on agarose-gel electrophoresis was treated as the true conditions, and the prediction by RNN was treated as the predictive conditions. A true positive, false negative, false positive, true negative, sensitivity, specificity, and accuracy were calculated accordingly. Significant differences in sensitivity, specificity, and accuracy between conditions were made based on Student’s and Welch’s t-test.

Results

PCR results of the primer sets and templates. PCR amplification with 72-sets of primer pairs on 31-templates was shown as 0 (no PCR amplification) or 1 (PCR product is visible) (Table 4). On 12 sets of primer pairs (Numbers 1, 2, 5, 17, 23, 25, 28–30, 41, 43 and 50), a PCR product was visible with more than 22 templates (Numbers 6, 8, 10, 11, 14, 16, 18, 19, 34–40, 48, 57, 58, 61, 63, 65 and 69). On the other hand, on 10 primer pairs (Numbers 7, 20, 21, 22, 46, 49, 55, 60, 68 and 71), no PCR product was shown. On 12 sets of primer pairs (Numbers 1, 2, 5, 17, 23, 25, 28–30, 41, 43 and 50), a PCR product was visible with more than 22 templates (Numbers 6, 8, 10, 11, 14, 16, 18, 19, 34–40, 48, 57, 58, 61, 63, 65 and 69). On the other hand, on 10 primer pairs (Numbers 7, 20, 21, 22, 46, 49, 55, 60, 68 and 71), no PCR product was shown. On the learning of the results, pseudo-sentences were created on each primer pair-template set. The pseudo-sentences were firstly classified by template and then randomly grouped into 5 groups to suppress the influences of a particular primer pair.

PCR prediction by recurrent neural network (RNN). As can be gleaned from the PCR results, the whole combination of primer pair and templates were divided into 5 groups (RNN-learned PCR results of 4 in 5 groups). Prediction accuracy on a verification group after learning on 4 groups was a plot against epochs.
Suitable for predicting PCR-positive results. On the other hand, an increase in the possibility of detecting PCR-negative sets is large in RNN learning. In the test sets, the particularly low sensitivity is seen in prediction in the whole data. These results suggest that the prediction rate of PCR-positives decreases when the number of test, the sensitivity was significantly higher in the undersampling data, and the specificity was significantly higher in comparison with whole-data and undersampling predictions, the sensitivity was significantly higher with undersampling data were lower than those of whole data. Similarly, in the test, the sensitivity was significantly higher in the undersampling data, and the specificity was significantly higher in the whole data. These results suggest that the prediction rate of PCR-positives decreases when the number of PCR negative sets is large in RNN learning. In the test sets, the particularly low sensitivity is seen in prediction.

In the division verification, sensitivity, specificity, and accuracy were calculated for the validation data in each division group. On the test data, the prediction was performed by learned RNN on each division verification. Thus, there were 5 predictions and sensitivity, specificity, and accuracy on test data. As a result, 5 sets of data were obtained under each condition, and a significant difference could be detected between those data (Table 6). In comparing whole-data and undersampling predictions, the sensitivity was significantly higher with undersampling. Moreover, in the specificity of undersampling, data were lower than those of whole data. Similarly, in the test, the sensitivity was significantly higher in the undersampling data, and the specificity was significantly higher in the whole data. These results suggest that the prediction rate of PCR-positives decreases when the number of PCR negative sets is large in RNN learning. In the test sets, the particularly low sensitivity is seen in prediction using whole-data-learned RNN (Table 5C). These results suggest that the current prediction method may depend on the base sequence of the primer itself. Thus, the coding method described in this study may not be perfectly suitable for predicting PCR-positive results. On the other hand, an increase in the possibility of detecting PCR positive through undersampling-data-learned RNN has indicated that prediction also depends on the number of negative and positive samples during learning (Table 5D).

Color summarization of prediction and result. To depict how PCR was predicted with individual primer-template combinations, we colored and displayed the individual PCR results as shown in Table 3 concerning the PCR-result and predictions in validation-data (Table 7). In this color display, no template strongly affected the prediction. On the other hand, several primer pairs suggested affecting PCR prediction (primer set numbers 17, 23, 41, and 43) (Table 7A). On the other hand, primer pair numbers 5, 24, 26, 50 and 70 showed relatively low accuracies through the templates. For this primer pair, it is suggested that RNN did not use much of this primer set information when making a prediction.
A. Prepare primers and primer set for each template.

| Primers and sequences | Template sequences |
|-----------------------|-------------------|
|                       |                   |

B. Search extendable hairpins and predict extended sequence for each primer.

```
5' CTTGGACATGCCTTGCTG 3'
5' CTTGGACATGCCTTGCTG 3'
5' CTTGGACATGCCTTGCTG 3'
```

Extended primers are added to each primer pair (*_hp).

C. Search extendable dimers and predict extended sequences for each primer set.

```
TTTGGACATGCCTTGCTG
CTTGACATGCCTTGCTG
CTTGACATGCCTTGCTG
AAAAGGGAGGCTTGCTG
```

D. Search homologous regions between primers and the template for each primer pair region which length is longer than 6 and number of mismatched bases is fewer than Table 4.

```
5' - CCATGCGAAAGGCTTGCTC-3' 3' - ATCCGACAA-5'
5' - CCATGCGAAAGGCTTGCTC-3' 3' - ATCCGACAA-5'
5' - CCATGCGAAAGGCTTGCTC-3' 3' - ATCCGACAA-5'
5' - CCATGCGAAAGGCTTGCTC-3' 3' - ATCCGACAA-5'
5' - CCATGCGAAAGGCTTGCTC-3' 3' - ATCCGACAA-5'
```

E. Assign pseudo-letters for each selected prime-template homologous regions. The pseudo-letters were determined referring to Table 4.

```
5' - CCATGCGAAAGGCTTGCTC-3' 3' - ATCCGACAA-5'
5' - CCATGCGAAAGGCTTGCTC-3' 3' - ATCCGACAA-5'
```

Outline of pseudo-letters.
F. Calculate predicting Gibbs energy on each primer-template homologous region. The Gibbs energy for each homologous region are calculated from divided dimers and predicted dimer-Enthalpy and Entropy (Supplement Table 4).

```
< Divide into Serial dimers and calculate ΔH and ΔS>

| Primer | Template |
|--------|----------|
| 5’-ATGCTGCGCC-3’ | 3’-TACTAACGCGCC-5’ |

Enthalpy: -72.4 -8.4 23.8 18.4 7.6 -8.4 -9.8 -10.6 -8.8 -8.0 -31.9 kcal/mol
Entropy: -20.4 -22.4 5.4 47.6 20.2 -22.4 -24.4 -27.2 -24.4 -19.9 -86.9 cal/mol

\[
T (\text{Absolute temperature}) = C (\text{Celsius temperature}) \times 270.15
\]
\[
\Delta G = (\Delta H - T \cdot \Delta S)
\]
\[
= (-31.9 - 329.15 \cdot (-0.0869))
\]
\[
= -3.29 \text{ kcal}
\]

G. Select priming sites on each primer pair-template and capitalize the pseudo-letters on each homologous region which Gibbs energy is smallest in each direction.

```
(8/8) 3’-GGCCCTAA-5’
5’-CCATCCGCAAGGCCTTTCCGGGATTCC-3’
```

```
(6/7) 3’-ATCCGAAA-5’
5’-CCATCCGCAAGGCCTTTCCGGGATTCC-3’
```

```
(8/8) 3’-ATCCGAAA-5’
5’-CCATCCGCAAGGCCTTTCCGGGATTCC-3’
```

H. Search hairpin and dimers on each extended primer pairs and assign pseudo-letters for each hairpin or dimer.

```
Hairpins

KLNNLLKK

Dimers

klklk lllklklll
```

Figure 2. (continued)
1. Create penta-codes and repeat according to the length of homologous region.
   a. The repeat number of primer-template penta-codes are determined according to base-distance from 3’-end of the homologous region (Table 6).

   \[
   \begin{align*}
   &\text{(8/8)’-CCATCGCAAGG} \quad \text{G6G6G6F} \\
   &\text{5’-CCATCGCAAGG} \quad \text{TTCCCCTGATCC-3’} \\
   &\text{[8-base homology \rightarrow 2,1,1]} \\
   &\text{G6G6G6, G6G6G6, F6G6G6, F6G6G6, FFG6G6, FFG6G6} \\
   &\text{5’-ATCCGCA-3’ (6/7) } \\
   &\text{5’-CCATCGCAAGG} \quad \text{TTCCCCTGATCC-3’} \\
   &\text{[7-base homology \rightarrow 2,1,1]} \\
   &\text{ttggg, ttggg, ttggg, ttggg} \\
   &\text{5’-ATCCGCA-3’ (8/8) } \\
   &\text{3’-GCTAGGGCTTCCGGAAAGGCCTAAGG-5’} \\
   &\text{AABBBAAA} \\
   &\text{[8-base homology \rightarrow 2,2,1,1]} \\
   &\text{BBBAA, BBBAA, BBBAA, BBBAA, AABBBA, AABBBA, AABBBA} \\
   \end{align*}
   \]

b. The repeat number of hairpin and dimer penta-codes are determined according to the base-length of hairpin stem or dimer-region (Table 7).

   \[
   \begin{align*}
   &\text{5’-TTTGGGATTAGCATACG} \quad G \quad C \\
   &\text{3’-AAACCTAATGCTATCC} \quad C \quad C \quad A \\
   &\text{KKKLLKKKLLKKKLLKKL} \\
   &\text{KKKLLKKKLLKKL} \\
   &\text{KKKLLKLLKKKLLKKL} \\
   &\text{KKKLLKLLKKKLLKKL} \\
   \end{align*}
   \]

J. Assign pseudo-letters to entire primers that suggested to prime PCR

Forward: ATCCCGGTTATCGGGCAA

\[
\text{Forward: PPPPPP PPPPPP PPPPPP PPPPPP} \\
\text{Reverse: GGGCCTAATGTGCCTGTTT} \\
\text{vwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww

K. Assemble pentacodes on each primer pair-template in the order of hairpin, dimer and positions of the homologous region on the template

**Figure 2. (continued)**
Table 4. PCR results with a combination of 72 primer sets and 31 templates. PCR results are expressed by the following numerical values based on observation on agarose gels. Number 0: No PCR product, 1: PCR product can be confirmed. Results on the primer pairs which have positive results on more than 20 templates are shown with pink background. Similarly, primer pairs which have results with single or no positive on 31 templates are shown with lime or blue background, respectively.
In undersampling data, two-thirds of the negatives were excluded from the prediction, so white squares are shown (Table 7B). For primer pairs that were PCR positive for many templates in this group, RNNs were often predicted to be correctly positive. From this result, it was suggested that RNNs whose degree of positive learning was increased by undersampling.

**Gibbs scatter plot.** When we created the pseudo-word, binding to template at the 3′ end of the primer sequence was recorded as the binding of the primer that could develop into PCR and used for prediction. At that time, the Gibbs energy was calculated for most of the bonds to the template at the 3′ end, and the primer at the position where the Gibbs energy was low and the PCR product was produced was used as a predictive primer-position for PCR prediction. Using this result, a scatter plot was created for forward and reverse with the assumed primer binding Gibbs energy on the horizontal and vertical axes (Fig. 5). When the set data with a positive PCR result is marked with a red triangle and those with a negative PCR result is marked with a blue circle, the set with the full length of the primer homologous to the template plot in the lower left. The set with the only weak binding plot in the upper right displayed (Fig. 5A). We predicted that PCR would occur only with strong interactions in the lower left region and not with weak interactions in the upper right region. While, in our PCR experiments, many PCR positives were found in the upper right region.

We also showed the prediction results were superimposed the Gibbs plot on the PCR-results (Fig. 5B,C). The primer pair-template sets for which the PCR-positive-RNN-negative result shown in pink were found not only in the upper right region but also in the lower left region in a similar ratio. If the prediction is based on the strength of hydrogen bonds, the prediction accuracy in the lower left region is expected to be high, but the prediction in this study did not show such a tendency. Even for the undersampling data with improved PCR-positive prediction accuracy, no improvement in prediction accuracy was observed in the lower left region (Fig. 5C).

Regarding the prediction of PCR-negative, many prediction errors related to PCR negative were observed in the peripheral region where the Gibbs energy of sets were slightly lower than those in the most upper right region (Fig. 5B). This trend did not change with the undersampling data (Fig. 5C). We expected that the PCR-negative prediction would use the same mechanism as the PCR-positive prediction, but in the lower left region of the whole sample prediction, the RNN correctly predicted several PCR-negatives in the lower left region (Fig. 5B). At the time of undersampling, these sets were not selected on the random selection and did not plotted (Fig. 5C). Considering that the prediction accuracy for PCR-positive in this region was not high, it is suggested that the PCR-negative prediction in the lower left region uses different criteria from the positive prediction.

**Discussion**

PCR is one of the basic technologies commonly utilized for genetic as well as pathogen-detection testing. Because of its declining cost, determining the base sequence of DNA or RNA subjected to PCR has now considerably increased. Furthermore, the development of applied technologies such as real-time and droplet PCRs and the application range of PCR has been expanded even further. When PCR is used to detect pathogens, specific detection is required. Moreover, such consideration for specific detection can be affected by some base sequence contaminants in processed samples. It is expected that such cases will likely increase if not rectified.

One of the strengths of PCR is that once a DNA is known, a highly sensitive test or method can be developed. This can be applied to various test targets over a very short period. It means that a highly sensitive detection becomes possible in a shorter time compared with developing an immunological test or technique. The only disadvantage of PCR method is, when there is a similar sequence between them, there is always a possibility that...
non-specific bands may be generated. This can happen as in the case of distinguishing bacteria by targeting a specific molecule that is contained only in ribosomal RNA. In this instance, it is difficult to design primers that enable specific detection because 16S ribosomal RNAs can have similar base sequences with each other. Thus, a test is required in the presence of a similar nucleotide sequence such as when a specific pathogen is found in a sample in which many other species are mixed.

Major parts of PCR primer design technologies were almost completed in the 1990s. The primer design technique is based on the stability of the hydrogen bond between the primer and the template based on the nucleotide sequence, and the PCR experiments conducted while examining its stability. Its hydrogen bond stability can be predicted by the free energy calculated from enthalpy, entropy, and absolute temperature. Early basic experiments have proven that one base at 3′ greatly affects the PCR reaction, and primers are designed based on 3′ several bases. Software for verifying the easiness of primer application and for designing primers by extracting the susceptible base sequence from the target base sequence was also developed. This primer design software, especially the Primer3, has a very large tracking record. Primers designed with Primer3 can amplify target DNA with an accuracy of 80% to 90%. However, even if the conventional primer design algorithm can design the primer that is most likely to cause PCR in the target template, it does not predict the amplification in the template DNA other than the intended one present in the sample. In our preliminary experiment, several Primer3-designed primer pairs amplify all 16sRNA templates regardless of the target DNA on the design of primers. Therefore, for a design of a primer pair that amplifies only the target template with the existence of similar sequences, it is necessary to consider a method different from the previous optimum design.

In the design of PCR primers, it is difficult to compare primer sets relative to each other by the method of selecting the optimum set. When selecting the optimum primer with Primer3 etc., 30 or more indexes are

Figure 4. Average prediction accuracy on validation groups in cross-validation. Average of prediction accuracy was calculated on 5 validation groups in cross-validation. (A) Whole sets in 72-primer-31template sets were used for learning or validation. (B) The number of primer pair-template sets in 72-primer-31template was controlled to 1:1 by undersampling. Groups are Orange: PCR positive primer pair-template sets, Green: PCR negative primer pair-template sets, and Blue: all primer pair-template sets. Standard deviation within validation groups was shown as error bars.
calculated, but a formula that uniformly shows the relationship between those indexes and the actual PCR is not provided\(^7\). It is expected that various DNAs in PCR tubes, including templates and primers, and PCR reaction conditions will contribute to the PCR results in different proportions under each condition. For example, the 3′ end of the primer is known to have a very large effect on PCR with just a few bases. Although DOP-PCR and similar arbitrary methods are known to nonspecifically amplify a wide range of DNA by matching several bases\(^44\). Experiments in the artificial gene synthesis from oligomers have also suggested that the primers are easily elongated when they form dimers\(^21\). Not only the binding position of the primer but also the base sequence of the PCR target region may have an influence depending on the annealing temperature. Of course, the base sequences of the primers and templates themselves also affect the results as factors other than mere stability. Thus, to design a primer that performs PCR only on a specific template, not on similar template sequences, it is necessary to consider the unknown number of factors without information about any contribution.

In recent years, supervised machine learning\(^14\) has been developed as a method of making predictions without determining the number or combination of factors that contribute to the results. In this method, after preparing data with correct answers, a large number of perceptrons are connected (perceptron network), and the serial adjustment of connection is optimized to form the perceptron network with the highest accuracy rate\(^14–17\). Since the substance of the prediction is a set of coefficients of the perceptron and its network, it is not necessary to analyze the factors for increasing the accuracy rate. Instead, analyzing learned machine learning often does not reveal the factors. Based on the discussion in the previous paragraph, it was expected that supervised machine learning would be suitable, as it does not require the number or combination of factors that contribute to the results to predict the success or failure of PCR.

In this study, PCR results were predicted from the base sequences of primers and templates using natural language processing that examines text trends. The PCR reaction is affected not by the base sequence of the primer or the template alone, but by the combination of complementary strands when they form a complementary strand. Therefore, we decided to generate the code from a combination of PCR primer pairs and complementary strand

| Prediction on RNN | PCR-result |  |
|-------------------|------------|--|
|                   | Negative   | Positive  |
| A. Validation of whole data | 1481 | 242 |
| Positive          | 197        | 312       |
| B. Validation of undersampling data | 380 | 174 |
| Positive          | 138        | 416       |
| C. Cross table on the test data predicted by whole data-learned RNN | 1419 | 73 |
| Positive          | 176        | 6         |
| D. Cross table on the test data predicted by whole data-learned RNN | 1051 | 42 |
| Positive          | 544        | 37        |

| Table 5. | Cross table of PCR results and predictions. Cross tables on PCR-results and RNN-predictions are made on the RNN-predictions on 200-epochs. Set numbers on RNN-prediction of the test data (C, D) were shown on each learned group on which prediction algorithm was created by learning. Data from prediction on group 1 is shown on C and D, average and standard deviations on prediction from 5 groups are listed in Table 6. |

| Validation  | Sensitivity | Specificity | Accuracy |
|-------------|-------------|-------------|----------|
| Whole data  | 0.563 ± 0.062 | 0.882 ± 0.024\(^4\) | 0.803 ± 0.029\(^1\) |
| Undersampling | 0.751 ± 0.031\(^2\) | 0.686 ± 0.021\(^7\) | 0.718 ± 0.022\(^2\) |
| Test        | 0.114 ± 0.092 | 0.899 ± 0.022\(^1\) | 0.862 ± 0.020\(^4\) |
| Undersampling | 0.471 ± 0.092 | 0.661 ± 0.056\(^4\) | 0.652 ± 0.052\(^2\) |

| Table 6. | Sensitivity, specificity and accuracy in split verification data and test data by learned RNN. Sensitivity, specificity and accuracy were calculated for RNN predictions and PCR results from 5 split validation groups for the whole data and undersampling sets respectively. For the Test data, the mean value and standard deviation were calculated for the results of prediction by the RNN that independently learned with each verification group. Numbers on superscript show groups in which no significant difference are detected. On the other combinations, significant differences were shown with the Student’s and Welch’s t-tests. |

Calculated, but a formula that uniformly shows the relationship between those indexes and the actual PCR is not provided\(^7\). It is expected that various DNAs in PCR tubes, including templates and primers, and PCR reaction conditions will contribute to the PCR results in different proportions under each condition. For example, the 3′ end of the primer is known to have a very large effect on PCR with just a few bases. Although DOP-PCR and similar arbitrary methods are known to nonspecifically amplify a wide range of DNA by matching several bases\(^44\). Experiments in the artificial gene synthesis from oligomers have also suggested that the primers are easily elongated when they form dimers\(^21\). Not only the binding position of the primer but also the base sequence of the PCR target region may have an influence depending on the annealing temperature. Of course, the base sequences of the primers and templates themselves also affect the results as factors other than mere stability. Thus, to design a primer that performs PCR only on a specific template, not on similar template sequences, it is necessary to consider the unknown number of factors without information about any contribution. In recent years, supervised machine learning\(^14\) has been developed as a method of making predictions without determining the number or combination of factors that contribute to the results. In this method, after preparing data with correct answers, a large number of perceptrons are connected (perceptron network), and the serial adjustment of connection is optimized to form the perceptron network with the highest accuracy rate\(^14–17\). Since the substance of the prediction is a set of coefficients of the perceptron and its network, it is not necessary to analyze the factors for increasing the accuracy rate. Instead, analyzing learned machine learning often does not reveal the factors. Based on the discussion in the previous paragraph, it was expected that supervised machine learning would be suitable, as it does not require the number or combination of factors that contribute to the results to predict the success or failure of PCR.

In this study, PCR results were predicted from the base sequences of primers and templates using natural language processing that examines text trends. The PCR reaction is affected not by the base sequence of the primer or the template alone, but by the combination of complementary strands when they form a complementary strand. Therefore, we decided to generate the code from a combination of PCR primer pairs and complementary strand
Table 7. (continued)
Table 7. Color summarization of PCR results and predictions, (A) predictions from whole-data, (B) predictions from undersampling data. Color presentations for PCR-result and RNN-prediction. Colors show the following results and predictions; red: PCR-positive-prediction-positive, pink: PCR-positive-prediction-negative, blue: PCR-negative-prediction-negative, light blue: PCR-negative-prediction-positive, white: primer pair-template sets excluded from predictions during undersampling.
Figure 5. Scatter plot of PCR-results and predictions. Plot the PCR results and RNN-predictions against the Gibbs energy at the hydrogen bond at the forward (horizontal axis) and reverse (vertical axis) priming positions determined at the time of pseudo-sentence determination (Fig. 2). Colors and shape show the following results and predictions; red triangle: PCR-positive-prediction-positive, pink triangle: PCR-positive-prediction-negative, blue circle: PCR-negative-prediction-negative, light blue circle: PCR-negative-prediction-positive. Primer pair-template sets excluded from predictions during undersampling were not plotted.
bases formed by the template. The generated code was split into words so that a sentence was formed from a set of primers and templates. Since a sentence can be created for each primer set and template, if there is a PCR experimental result, the PCR experimental result can be linked as a correct answer to each sentence. In natural language processing, a machine-learning network is made to learn a sentence whose evaluation is confirmed, and the learned network predicts an unidentified sentence. In RNN, which is a typical natural language processing machine learning, RNN is trained in movie criticisms with positive evaluation and negative movie criticism, and the evaluation is predicted for unidentified movie criticisms. By generating pseudo-sentences using the primer set and template proposed in this paper as a unit, it is possible to associate PCR results with each pseudo-sentence in the same way as Positive/Negative in film criticism. Since the generation of pseudo-words from the complement set alone could not reflect that the complementarity of the 3′end was greater than that of the 5′end, it was emphasized as a word iteration. Therefore, for the learning of pseudo-sentences in this study, the same RNN as the one learned for the evaluation of film criticism was used. This is the first paper to use a neural network application to design primers and predict PCR results. Supervised machine learning was used to learn the PCR results. Since we created pseudo-words and pseudo-sentences as input information, we selected RNNs to learn the relationship between primer and template sequences and PCR results. RNNs can interpret sentences while analyzing the context of words in the sentence. In this study, in a test experiment conducted by actually creating a new primer, prediction was made with an accuracy of 70% or more (Table 5). These results suggest that the interaction between the primer and the template is also effective when the interpreted data of the RNN is returned to the previous layer and used for further interpretation. They also suggest that the effect of primer-template interaction on PCR is similar to the effect of natural language word placement in semantic interpretation. The LSTM used the word context in the sentence to change the retention of the word's effect for each word and make a comprehensive judgment of its effectiveness. We created our pseudo-words for RNN analysis for this study (Fig. 2). All of the letters that make up a word were determined based on the primer-template interactions that are important in previous studies (Fig. 2E). Natural language processing by RNN uses all the words used in a specific language, so the vocabulary is about 30 to 100,000 words (RNN literature). In this study, the data was as small as 2,000, so it was necessary to have a small vocabulary. Therefore, the original 16 base combinations are summarized in 5 based on the effect of Taq polymerase on DNA synthesis. However, considering that the primers face each other in the opposite direction during PCR, the direction of homology was reflected in the letters. Besides, different character sets were prepared for dimers and hairpins. Also, uppercase and lowercase letters were set for the evaluation target as the starting point of PCR and other parts. As a result, the vocabulary of the 5-letter pseudoword (pentacode) code was 5 to the 5th power × 5 × 2 = 31,250. In RNN, the characteristics of each sentence are expressed by the amount of words used (word vector) with the vocabulary as the number of dimensions. If the vocabulary is large, the frequency of occurrence of words is low, so the word vector becomes a sparse vector and may not sufficiently show the characteristics of the sentence. On the other hand, when the size of the vocabulary is small, detailed features may not be expressed, which suggests that the prediction accuracy is limited. In the method of this study, the number of characters in a word was shortened to 5 as another method to reduce the size of the vocabulary. It is suggested that extending this to 6 or 7 bases will increase the vocabulary and enable more accurate predictions. In the future, it is thought that this code setting method can be improved by accumulating more data.

In this study, pseudo-words were created based on primer hairpins, dimers, primer-template homology, and primer-PCR product homology. Predicting the priming position is expected to be particularly important among pseudo-words. This is because PCR is established based on the elongation of DNA from the priming position (Fig. 1). When designing the optimum primer as in the conventional case, the binding position of the primer has a long complementary region and high stability as compared with other positions. However, when comparing the complementarity between the template and the primer sequence, which was not originally designed, it is necessary to determine the priming position from a large number of candidates having similar length and stability of the complementary strand. Also, the effect of priming position was conveyed by expressing the priming position in capital letters. The accuracy of this pumping position affects the accuracy of the overall prediction, whereas, in addition to the complementary with the base sequence and template of the primer, it becomes an amplified sequence or set (reverse for forward, forward for reverse). Thus, its relationship with the priming position is also affected. Therefore, it is ideally desirable to learn and predict this priming position by artificial intelligence. However, since the basic data is not available in this study, the stability of the complementary strand is predicted by the nearest neighbor method. The priming position that maximizes stability was predicted with the set of priming positions. For the prediction of free energy by the nearest neighbor method, in addition to the values reported so far, values extrapolated from those values were set and used. Since some of these numbers are simple extrapolations from the reported numbers, their accuracy is not yet guaranteed, hence, future improvements are still needed.

Improvement in prediction accuracy in RNN is enhanced in the process of repeating epochs (Fig. 4). When all the data were used, the prediction was stable at about 25 epochs, and no significant change occurred. After which, when the number of PCR positive and negative data was matched by undersampling, the error was up to 75 epochs larger. Later transition period of up to 100 epochs made the prediction accuracy become stable. This indicates that the structure of data affects the learning steps of RNN. When the number of data or composition is changed in the future, we proposed to first investigate the changes in epoch and prediction accuracy.

The PCR results used in this study include those that were greatly influenced by primer pairs (Table 4). In 12 of 72 primer pairs, PCR was observed in 20 or more of 31 templates. In 22 primer pairs, PCR was observed in only 1 template. No PCR was observed for 10 primer pairs. Perspectives, these primer pair-template data combinations showed that the predictions were relatively correct when only one of the templates was amplified or when PCR was not applied to any of the templates (Table 7A). This suggests that PCR was successful to primers with high specificity, and conversely, RNNs made highly accurate predictions for primer sets with low PCR characteristics.
On the other hand, in the primer pairs in which PCR was observed in a large number of templates, the prediction was relatively wrong, suggesting that it was difficult to predict the RNN for such primers in where false positives frequently appear. The relationship between primer binding to the template and prediction is shown in a scatter plot made with Gibbs energy at the optimal binding position of the primer (Fig. 5). In this scatter plot, the primer and template set specifically designed for lower left lower region are shown, and the results for the primer pair and template set that do not assume PCR are shown in the upper right region (Fig. 5A). Surprisingly, the prediction did not always hit lower left region, but to the same extent in the upper right (Fig. 5B,C). This tendency was the same for undersampling, suggesting that improvement in prediction accuracy for PCR positive was influenced by improvement in the accuracy rate in the upper right region. For PCR-negative predictions, it is noteworthy that the RNN hit the predictions for multiple PCR-negative sets in the lower left region of the scatter plot created from the predictions of all the data. These results show that the RNN described in this study does not have high accuracy at present, but it is expected that the prediction accuracy will be improved by improving the number of data and reviewing pseudo-words in the future.

It is challenging for RNNs to simplify which of the pseudo-words and their repetitions can have a great influence on the characteristics of supervised machine learning. The correctness of the prediction does not guarantee the correctness of the setting like the pseudo-word. Moreover, through this paper, researchers may now find it useful to reconstruct the prediction method. Pseudo-word generation and pseudo-sentence prediction do not provide the theoretical justification of algorithms based on unified theory, but databased reproducibility can be provided to the user.

In conclusion, it is indicated that PCR design by natural language processing system using RNN be utilized in enabling a primer design to detect a specific template in the presence of multiple templates. Method accuracy is improved by learning the base sequence of the primer pair, the template, and the PCR result. Design can be upgraded by using discarded negative data.

Received: 4 May 2020; Accepted: 9 March 2021
Published online: 05 April 2021

References
1. Dieffenbach, C. W., Lowe, T. M. & Dveksler, G. S. General concepts for PCR primer design. PCR Methods Appl. 3(3), S30–S37. https://doi.org/10.1101/gr.3.3.s30 (1993).
2. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. Bioinformatics 23(10), 1289–1291. https://doi.org/10.1093/bioinformatics/btm091 (2007).
3. Rychlik, W., Spencer, W. I. & Rhoads, R. E. Optimization of the annealing temperature for DNA amplification in vitro [published correction appears in Nucleic Acids Res 1991 19(3), 698]. Nucleic Acids Res. 18(21), 6409–6412. https://doi.org/10.1093/nar/18.21.6409 (1990).
4. Graham, K. J. & Holland, M. J. PrimerSelect: A transcriptome-wide oligonucleotide primer pair design program for kinetic RT-PCR-based transcript profiling. Methods Enzymol. 395, 544–553. https://doi.org/10.1016/S0076-6879(05)95028-3 (2005).
5. Rychlik, W. OLG07 primer analysis software. Methods Mol. Biol. 402, 35–60. https://doi.org/10.1007/978-1-59745-528-2_2 (2007).
6. SantaLucia, J. Jr. Physical principles and visual-OMP software for optimal PCR design. Methods Mol. Biol. 402, 3–34. https://doi.org/10.1007/978-1-59745-528-2_1 (2007).
7. Undergasser, A. et al. Primer3—New capabilities and interfaces. Nucleic Acids Res. 40(15), e115. https://doi.org/10.1093/nar/gks596 (2012).
8. SantaLucia, J. Jr. & Hicks, D. The thermodynamics of DNA structural motifs. Annu. Rev. Biophys. Biomol. Struct. 33, 415–440. https://doi.org/10.1146/annurev.biophys.32.110601.141800 (2004).
9. Andreason, R., Repp, E., Kaplinski, L. & Remm, M. GENOMEMASKER package for designing unique genomic PCR primers. BMC Bioinform. 7, 172. https://doi.org/10.1186/1471-2105-7-431 (2006).
10. Pecchia, S. et al. Molecular detection of the seed-borne pathogen Colletotrichum lundii targeting the hyper-variable IGS region of the ribosomal cluster. Plants 8(7), 222. https://doi.org/10.3390/plants8070222 (2019).
11. Gorny, A. M., Wang, X., Hay, F. S. & Pethybridge, S. J. Development of a species-specific PCR for detection and quantification of meloidogyne hapla in soil using the 16D10 root-knot nematode effector gene. Plant Dis 103(8), 1902–1909. https://doi.org/10.1094/PDIS-09-18-1539-R (2019).
12. Chen, Q., Jun, L., Qiu, Y. & Zhao, L. Short communication: Bioinformatics-based mining of novel gene targets for identification of Cronobacter tuevisensis using PCR. J. Dairy Sci. 102(7), 6023–6026. https://doi.org/10.3168/jds.2018-15929 (2019).
13. Afin, K. S. et al. Development of molecular marker through genome realignment for specific detection of Xanthomonas campestris pv. campestris Race 5, a pathogen of black rot disease. J. Microbiol. Biotechnol. 29(5), 785–793. https://doi.org/10.4014/jmb.1901.01050 (2019).
14. Kotsiantis, S. B., Zaharakis, I. & Pintelas, P. Supervised machine learning: A review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. 160, 3–24 (2007).
15. Tsai, C. F. et al. Intrusion detection by machine learning: A review. Expert Syst. Appl. 36(10), 11994–12000. https://doi.org/10.1016/j.eswa.2009.05.029 (2009).
16. Pang, B., Lee, L., & Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing—Volume 10 (EMNLP ’02). Association for Computational Linguistics, USA, 79–86 https://doi.org/10.3115/1118693.1118704 (2002).
17. Sebastián, F. Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47. https://doi.org/10.1145/505282.505283 (2002).
18. Banko, M., & Brill, E. Scaling to very very large corpora for natural language disambiguation. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL ’01). Association for Computational Linguistics, USA, 26–33 https://doi.org/10.3115/1073012.1073017 (2001).
19. Lewis, D. Naïve (Bayes) at forty: The independence assumption in information retrieval. In ECML: Machine Learning. 4–15 https://doi.org/10.1007/BFb0062666. (1998).
20. Mikolov, T., et al. Recurrent neural network based language model. In Eleventh Annual Conference of the International Speech Communication Association (2010). https://www.isca-speech.org/archive/interspeech_2010/I10_1045.html.
21. Kayama, K., Hashizume, H., Camer, G.A. & Endoh, D. An improved gene synthesis method with asymmetric directions of oligonucleotides designed using a simulation program. *Biotechniques* **69**(3), 211–219. https://doi.org/10.2144/btn-2020-0062 (2020).

22. Xiao, Y.H. & Pei, Y. Asymmetric overlap extension PCR method for site-directed mutagenesis. *Methods Mol. Biol.* **687**, 277–282. https://doi.org/10.1007/978-1-60761-944-4_20 (2011).

23. Horne, M.T., Fish, D.J. & Benight, A.S. Statistical thermodynamics and kinetics of DNA multiplex hybridization reactions. *Biophys. J.* **91**(11), 4133–4153. https://doi.org/10.1016/biophysj.106.090662 (2006).

24. De Winter, J.C.F. Using the Student’s t-test with extremely small sample sizes. *Pract. Assess. Res. Eval.* **18**(1), 10. https://doi.org/10.7275/e46d-d05 (2013).

25. Wang, R.F., Cao, W.W. & Cerniglia, C.E. PCR detection and quantitation of predominant anaerobic bacteria in human and animal fecal samples. *Appl. Environ. Microbiol.* **62**(4), 1242–1247 (1996).

26. Stadhouders, K. et al. The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5′-nuclease assay. *J. Mol. Diagn.* **12**(1), 109–117. https://doi.org/10.2355/jmoldx.2010.090035 (2010).

27. Lu, H., Giordano, F. & Ning, Z. Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinform.* **14**(5), 265–279. https://doi.org/10.4161/gpb.2016.05.004 (2016).

28. Levy, S.E. & Myers, R.M. Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.* **17**, 95–115. https://doi.org/10.1146/annurev-genom-083115-022413 (2016).

29. van Dijk, E.L., Auger, H., Jassyczyn, Y. & Therme, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**(9), 418–426. https://doi.org/10.1016/j.tig.2014.07.001 (2014).

30. Wei, Z. et al. Diagnostic accuracy of in-house real-time PCR assay for *Mycobacterium tuberculosis*: A systematic review and meta-analysis. *BMC Infect. Dis.* **19**(1), 701. https://doi.org/10.1186/s12879-019-4273-z (2019).

31. Lucas, E.J., Leber, A. & Ardura, M.I. Broad-range PCR application in a large academic pediatric center: Clinical value and challenges in diagnosis of infectious diseases. *Pediatr. Infect. Dis. J.* **38**(8), 786–790. https://doi.org/10.1097/INF.0000000000002308 (2019).

32. Tong, Y., Shen, S., Jiang, H. & Chen, Z. Application of digital PCR in detecting human diseases associated gene mutation. *Cell Physiol Biochem.* **43**(4), 1718–1730. https://doi.org/10.1159/000484035 (2017).

33. Arty, M. et al. Basic principles of real-time quantitative PCR. *Expert Rev. Mol. Diagn.* **5**(2), 209–219. https://doi.org/10.1586/14731585.5.2.209 (2005).

34. DiCesare, J. et al. A high-sensitivity electrochemiluminescence-based detection system for automated PCR product quantitation. *Biotechniques* **15**(1), 152–215 (1993).

35. Noordhoek, G.T. et al. Sensitivity and specificity of PCR for detection of *Mycobacterium tuberculosis*: A blind comparison study among seven laboratories. *J. Clin. Microbiol.* **32**(2), 277–284 (1994).

36. He, Z. et al. Improved PCR primers to amplify 16S rDNA genes from NC10 bacteria. *Appl. Microbiol. Biotechnol.* **100**(11), 5099–5108. https://doi.org/10.1007/s00253-016-7477-9 (2016).

37. Clifford, R. J. et al. Detection of bacterial 16S rDNA and identification of four clinically important bacteria by real-time PCR. *PLoS ONE* **7**(11), e48558. https://doi.org/10.1371/journal.pone.0048558 (2012).

38. Lin, C.K. et al. An improved PCR primer pair based on 16s rDNA for the specific detection of *Salmonella serovars* in food samples. *Foods* **6**(7), 1335–1343. https://doi.org/10.3390/foods6071335 (2017).

39. Baker, G.C., Smith, J.J. & Cowan, D.A. Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods.* **55**(3), 541–555. https://doi.org/10.1016/j.mimet.2003.08.009 (2003).

40. Annamalai, P., Ishii, H., Lalithakumari, D., & Revathi, R. Polymerase chain reaction and its applications in fungal disease diagnosis/ Polymerase-Kettenreaktion und ihre Anwendungen zur Diagnose von Pilzkrankheiten an Pflanzen. *Zeitschrift für Pflanzenkrankheiten und Pflanzenzüchtung.* **Plant Dis. Prot.* **91**–104 (1995).

41. Lomovzov, A.A., Vorobjov, Y.N. & Pyshnyi, D.V. Evaluation of the Gibbs free energy changes and melting temperatures of DNA/DNA duplexes using hybridization enthalpy calculated by molecular dynamics simulation. *J. Phys. Chem. B* **119**(49), 15221–15234 (2015).

42. Kwok, S. et al. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nucleic Acids Res.* **18**(4), 999–1005. https://doi.org/10.1093/nar/18.4.999 (1990).

43. Rozen, S. & Skalleyski, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 355–368. https://doi.org/10.1385/1-59259-192-2:365 (2000).

44. Teraoka, H. et al. Differential display system with vertebrate-common degenerate oligonucleotide primers: Uncovering genes responsive to dioxin in avian embryonic liver. *Environ. Sci. Technol.* **46**(1), 27–33. https://doi.org/10.1021/es201085m (2012).

45. Li, Y. et al. Sequential person recognition in photo albums with a recurrent network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).

46. Yin, W. et al. Comparative study of CNN and RNN for natural language processing. arXiv preprint, arXiv:1702.01923 (2017).

47. Wen, T.H. et al. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. arXiv preprint, arXiv:1508.01745 (2015).

48. Pan, X. et al. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* **19**, 511. https://doi.org/10.1186/s12864-018-4899-1 (2018).

Acknowledgements
Special thanks to Mr. Shiqiao Du for his advice on PyTorch and for the use of Python scripts that is modified from his book. This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT)-Supported Program for the Strategic Research Foundation at Private Universities, 2013–2017. Thanks is also due to the research funding of Rakuno Gakuen University with Lagrange Co. Ltd (Tokyo, Japan), 2018–2019.

Author contributions
K.K. and D.E. wrote the manuscript and prepared the figures. M.K., N.C., M.T., R.Y. and K.H. conducted the experiments. D.E. designed the algorithm and developed the software. G.A.C critically organized and enhanced the manuscript. All authors reviewed the manuscript.

Competing interests
K.K. and D.E. have applied for a patent in Japan for primer design using machine learning. All other authors have no competing interests.
