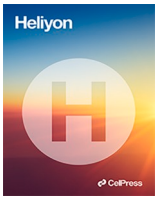
Heliyon 9 (2023) e18829

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



# Models for prediction of single base primer extension efficiency from position and type of single mismatch in primer-template duplex

Myong-Rim Ri, Jin-Sok Kang, Myong-Ryong Ri, Song Nam U *

*Department of Life Science, University of Science, Pyongyang, Democratic People's Republic of Korea*

ARTICLE INFO

*Keywords:*
Multiple linear regression models
Artificial neural network models
Single base extension

ABSTRACT

Amplification and specificity of polymerase chain reaction (PCR) are affected by the position and type of primer-template mismatches (MMs) as well as various conditions of reaction. In this study, multiple linear regression (MLR) models and artificial neural network (ANN) models were developed for the prediction of the effects of primer-template mismatches on the primer extension efficiency in primer-template duplex. In MLR models, the independent variable $P_i$ representing the position effect of $i$-th mismatch from 3′ end of primers was normalized to values between 0 and 1 according to the size of $\Delta\Delta G_i$, the difference of Gibbs free energy changes between the mismatch and its corresponding perfect-match, and other independent variables $P_j$ representing the position effect of the $j$-th perfect-match from 3' end of primer were coded 1. A dependent variable of MLR model was relative extension efficiencies of primers. In ANN models, an input layer has neurons equal to the number of independent variables of corresponding MLR models and a hidden layer and an output layer have four and one neurons, respectively. Our MLR models and ANN models outperform the previous polynomial regression model for the prediction of the single base extension (SBE) efficiencies of single-MM primers. Especially, ANN model 6 which has 32 neurons representing the position effect of mismatch, the type of mismatch and the annealing temperature on primer-template duplex in the input layer can predict the SBE efficiencies of single-MM primers with a high accuracy, since its correlation coefficients R in training set, testing set and all data are 0.9870, 0.9782 and 0.9857, respectively. These results will have a good prospect applicable to the design of primer and testing the primer specificity in genome database.

## 1. Introduction

Polymerase chain reaction (PCR) is a technology of molecular biology widely used for the isolation of gene, diagnosis of disease, identification of pathogens and forensic identification [1]. Amplification efficiency and specificity of PCR are affected by the position and type of primer-template mismatch as well as various conditions of reaction [2]. In particular, the specificity of primers is very significant when the infection by viruses with high mutation rate is tested with PCR [3].

Recently 19 different primer/probe sets (297 primers and 43 probes) for reverse transcription-polymerase chain reaction (RT-PCR) were designed from seven different genes of severe acute respiratory syndrome-coronavirus-2(SARS-COV-2) genomes (Nucleocapsid

* Corresponding author.
*E-mail address:* yurijingrixing@163.com (S.N. U).

https://doi.org/10.1016/j.heliyon.2023.e18829
Received 12 January 2023; Received in revised form 12 July 2023; Accepted 31 July 2023
Available online 1 August 2023

protein gene N, Envelope protein gene E, Spike protein gene S, Membrane protein gene M, Open Reading Frame ORF1ab, RNA-dependent RNA polymerase gene RdRp and Non-Structural Protein gene nsp2) by World Health Organization, Center of Disease Control and etc. and used for the diagnosis of COVID-19 and the discrimination of variants of concern (VOCs) [3]. Miranda and Weber [3] have showed that only some primer/probe sets align substantially to most SARS-CoV-2 genomes, and almost 75% of these sets also cross-align to some SARS-CoV-1 and non-SARS viruses, by aligning all primer/probe sets to 21,665 genomes of SARS-COV-2 and 323 genomes of other coronaviruses including SARS-COV-1 and calculating their melting temperatures taking account of up to three consecutive MMs. For the calculation of melting temperature of primer/probe sets, they have used $a_0, a_1$ regression coefficients obtained from 4,032 sequences containing up to three consecutive MM base pairs and their experimental melting temperatures and $\tau$, statistical index calculated from the classical partition function of a model Hamiltonian [3].

So far, several parameters based on the nearest-neighbor thermodynamics with zero $\Delta C_P$ [4–7] or non-zero $\Delta C_P$ [8,9] have been used for the calculation of melting temperature ($T_m$) of Primer/probe sets containing mismatches. Wu et al. [10] have systematically compared the relative extension efficiencies of perfect-matched (PM) and mismatched (MM) primers and showed that the relation between $\Delta T_m$, the differences of melting temperature in PM and MM primers determined using the DINAMelt web server and the relative extension efficiencies of MM primers can be represented as empirical polynomial equation and based on this relation, the relative extension efficiencies for better refining primer sequences and operational conditions can be roughly estimated.

Some authors [1,10–13] have shown that the efficiencies of PCR are dependent on the position and type of primer-template mismatches. Wu et al. [10] have early found that PCR or extension of primers are inhibited by single mismatches at the primer 3′ end and other positions and have continuously compared extension efficiencies between a PM primer and its single-MM primers with all possible MM types with a single base extension (SBE) assay. In SBE assay, after a tested primer that completely complementary to or has a single MM to the target deoxyribonucleic acid (DNA) sequence and a reference primer that differs from the tested primer in the target site and length are bound onto the targeted DNA sequences and extended with fluorescently labeled 2′,3′-dideoxyribonucleoside 5′-triphosphates (ddNTPs), extended primers can be separated by electrophoresis and SBE efficiencies can be determined from the fluorescence intensities of extended primers [10]. As a result, the fluorescence intensity ratio of PM and MM primers to the reference primer can be obtained and normalizing the fluorescence intensity ratio of PM primer to the reference primer equivalent to 100% efficiency, the relative extension efficiency of any given MM primer can be calculated accordingly [10]. SBE efficiencies of single-MM primers were generally lower or equivalent to that of PM primers, depending on the position and type of MMs, and MMs existing at the 1–4th positions from the primer 3′end inhibited strongly the SBE of primers but the internal MMs and MMs existing at the primer 5′end inhibited it weakly [10]. Bru et al. [1] studied the effects of the internal MMs in the primer-templates on the amplification efficiencies of PCR using the 16S rRNA genes as the template DNA and observed that the presence of MMs in the second half of the primer sequence can result in an underestimation of up to 1,000-fold of the gene copy number, depending on the primer and position of MMs. Kwok et al. [11] investigated the effects of various primer-template MMs on DNA amplification of an human immunodeficiency virus (HIV)-1 gag region by PCR and revealed that internal single MMs had no significant effect on PCR product yield, but MMs at the 3′-terminal base had different effects. In this case, 3′-terminal MMs such as A:G, G:A, and C:C reduced overall yield of PCR product about 100-fold, but all other 3′-terminal MM primers were efficiently amplified [11]. Double mismatches of which one is at the 3′ end of primer, in general, reduced PCR product yield dramatically, but even two MM Ts at the 3′ end of primer allowed the significant amplification [11]. Stadhouders et al. [13] studied the effect of primer-template MMs on real-time PCR using 5′-nuclease assay and demonstrated that single MMs such as A:C, C:A, T:G, G:T corresponds to cycle thresholds of $<1.5$, and MMs such as A:A, G:A, A:G, C:C corresponds to cycle thresholds of $>7.0$ and there exists a clear relationship between the types and position of MMs and the efficiencies of PCR. Ledeker and De Long [12] investigated an approach to quantify accurately pcrA, a gene encoding perchlorate reductase to elucidate quantitatively the effect of multiple primer–template MMs on quantitative PCR accuracy. They showed that for multiple MMs up to 3 MMs in the middle region and 5′ end of primer, quantification accuracies could be as low as ~0.1% [12]. Furthermore, when PCRs were implemented using a known pcrA primer pair with mixtures of genomic DNA from strains known to harbor the target gene, for some mixtures quantification accuracy was as low as ~0.8% or was not detected [12].

Although the sequence and length of primers, conditions of PCR and experimental approaches tested by some authors [1,10–13] were different, effects of the internal MMs and the 5′-end MMs of primers on the amplification efficiencies of PCR were relatively consistent, but the 3′-end MMs of primers had clear different impacts.

Miranda et al. [3] analyzed the coverage and cross-reactivity of available primers for SARS-CoV-2 and for non-SARS-CoV-2 viruses based on the difference between the calculated hybridization temperatures $T_{MM}$ taking into account of the mismatches and the reference temperature $T_{ref}$, hybridization temperature of the fully matched primer, but the polynomial equation showing the relation between melting temperatures of PM primers and MM primers and primer extension efficiencies proposed by Wu et al. [10], can only roughly estimate the relative extension efficiencies from primer sequences with the coefficient of determination $R^2 = 0.5318$.

Some authors [1,11–13] demonstrated that the type and position of primer-template MMs affected the amplification efficiencies of PCR, but did not show the quantitative relation between them.

Multiple linear regression (MLR) analysis is one of the most basic mathematical models. It is based on linear relationships with both inputs and outputs [14]. MLR is often used for the model development and the variable selection in model building [15]. Artificial intelligence such as an artificial neural network (ANN) is another method that can be used for the pattern regression and the model building [15]. ANN has the ability to derive meaningful relationships between imprecise data by linking input variables with each other and with the output values. The basic form of ANN consists of three layers: input, output, and hidden layer [14]. Among the various prediction methods, ANN gives a more accurate and efficient manner for predicting and analyzing the vast dataset [14]. But MLR and ANN models have not yet used for predicting SBE efficiencies of single MM primers.

From the discussion of the previous literature review, it is apparent that SBE efficiencies of single MM primers were generally lower

or equivalent to that of PM primers, depending on the position and type of MMs and the polynomial regression equation showing the relation between melting temperatures of MM primers and primer extension efficiencies proposed previously can only roughly estimate the SBE efficiencies from single MM primers. MLR and ANN Models considering directly the position and type of MMs in primer-template duplexes have not used in predicting SBE efficiencies of single MM primers. Elucidation of the quantitative relation between the primer extension efficiencies and the type and position of MMs can be a great help to the design of primers with high specificity and degenerate primers, and to the *in silico* test of the specificity and PCR efficiencies of primers based on genome sequence database. The major focus of this paper is to (1) select the necessary variables and the encoding method through the construction of MLR model for predicting SBE efficiencies of single MM primers from primer-template sequences. (2) To improve the prediction accuracy of SBE efficiencies of single MM primers using ANN model based on experimental data proposed by Wu et al. [10].

## 2. Materials and methods

### 2.1. Materials

Total 227 SBE efficiencies data obtained from 2 p.m. primers and 111 MM primers by SBE assay using CEQ™ single nucleotide polymorphism (SNP)-primer extension kit reported by Wu et al. [10] were used in MLR analysis and ANN training for the prediction of the SBE efficiencies from the position and type of single MMs. They include 114 SBE efficiencies data obtained at annealing temperature 53 °C and 60 °C, respectively using 16S-ITS-23S rRNA gene fragment of *Bacteroides coagulans* as template and 57 SBE efficiencies data obtained at annealing temperature 60 °C using 16S rRNA gene of *Stenotrophomonas acidaminiphila* as template about 57 all possible single MM primers designed from primer 1492R (5′-GGCTACCTTGTTACGACTT) that targets the conserved region of all prokaryotic 16S rRNA genes, 54 SBE efficiencies data obtained at annealing temperature 60 °C using 16S-ITS-23S rRNA gene fragment of *Methanosaeta concilii* as template about 54 all possible single MM primers designed from primer MX (5′- GCATCTCGACAGCCAGAT) that targets 16S rRNA gene of *Methanosaeta concilii* and 2 SBE efficiencies obtained from 2 p.m. primers. SBE efficiency is the relative extension efficiency of any given MM primer determined, by normalizing the fluorescence intensity ratio of PM primer to the reference primer equivalent to 100% efficiency.

### 2.2. Methods

#### 2.2.1. Construction of MLR models for the prediction of SBE efficiencies of single MM primers and regression analysis

For the prediction of SBE efficiencies of single MM primers, following 6 MLR models were constructed.

Regression model 1 was composed of 19 independent variables representing the position effect of MMs and the regression equation can be written as follows.

$$y = \beta_0 + \sum_{i=1}^{19} \beta_i x_i = \beta_0 + \sum_{i=1}^{19} \beta_{pi} P_i \tag{1}$$

Where $y$ is the dependent variable representing the SBE efficiencies of primers, $x_i$ denotes the $i$-th independent variable and $\beta_i$ is the population regression coefficient. $P_i$ is the independent variable representing the portion of Gibbs free energy that MM located at the $i$-th position from 3′ end of primer in primer-template duplex contributes to the duplex stability.

Regression model 2 was composed of 19 independent variables representing the position effect of MMs and an independent variable representing the annealing temperature, and the regression equation can be written as

$$y = \beta_0 + \sum_{i=1}^{20} \beta_i x_i = \beta_0 + \sum_{i=1}^{19} \beta_{pi} P_i + \beta_{Ann} Ann \tag{2}$$

where Ann is the independent variable representing the annealing temperature.

Regression model 3 was composed of 12 independent variables representing the type of MMs and the regression equation can be written as

$$y = \beta_0 + \sum_{i=1}^{12} \beta_i x_i = \beta_0 + \sum_{i=1}^{12} \beta_{MMi} MM_i \tag{3}$$

where $MM_i$ is the independent variable representing the $i$-th type of MMs existing in duplex.

Regression model 4 was composed of 12 independent variables representing the type of MMs and an independent variable representing the annealing temperature.

$$y = \beta_0 + \sum_{i=1}^{13} \beta_i x_i = \beta_0 + \sum_{i=1}^{12} \beta_{MM_i} MM_i + \beta_{Ann} Ann \tag{4}$$

Regression model 5 was composed of 19 independent variables representing the position effect of MMs and 12 independent variables representing the type of MMs.

$$y = \beta_0 + \sum_{i=1}^{31} \beta_i x_i = \beta_0 + \sum_{i=1}^{19} \beta_{P_i} P_i + \sum_{i=1}^{12} \beta_{MM_i} MM_i \tag{5}$$

Regression model 6 was composed of 19 independent variables representing the position effect of MMs, 12 independent variables representing the type of MMs and an independent variable representing the annealing temperature, and the regression equation can be written as

$$y = \beta_0 + \sum_{i=1}^{32}\beta_i x_i = \beta_0 + \sum_{i=1}^{19}\beta_{Pi}P_i + \sum_{i=1}^{12}\beta_{MMi}MMi + \beta_{Ann}Ann \quad (6)$$

The dependent variable representing the primer extension efficiency in all regression models was normalized to the value between 1 and 0, by dividing the relative extension efficiencies by 100.

In MLR models, 19 independent variable $P_i$ representing the position effect of $i$-th mismatch from 3' end of primers was normalized with formula 1 $-\Delta\Delta G_i/\Delta\Delta G_{max}$, where $\Delta\Delta G_i$ is the difference between the sum of $\Delta G^0_{37}s$ for two mismatched nearest-neighbor doublets $N_i{-}1N_i/C_{i\text{-}1}M_i$ and $C_{i+1}M_i/N_{i+1}N_i$ ($N_k$: a base of $k$-th position, $C_i$: a complementary base of $i$-th position, $M_i$: a mismatched base of $i$–th position) with the $i$-th mismatch as its axis and the sum of $\Delta G^0_{37}s$ for two corresponding complementary doublets $N_i{-}1N_i/C_{i\text{-}1}C_i$ and $C_{i+1}C_i/N_{i+1}N_i$, and $\Delta\Delta G_{max}$ is the maximum of possible $\Delta\Delta G_i$s. $\Delta G^0_{37}s$ for mismatched and complementary nearest-neighbor doublets were the nearest-neighbor parameters of SantaLucia [6,7]. All other independent variables $P_j$ representing the position effect of $j$-th perfect-matched base pairing were coded 1.

12 independent variables $MM_i$ representing the type of mismatches in order of AA, AC, AG, CA, CC, CT, GA, GG, GT, TC, TG and TT were coded either 1 or 0 to indicate presence or absence of types of MMs in primer-template duplexes.

An independent variable Ann representing annealing temperature ($T_{anneal}$) was normalized to values between 1 and 0 by the distribution $e^{-\left(\frac{T_{opt}-T_{anneal}}{\sigma}\right)^2}$ where the mean is the optimal annealing temperature of primer ($T_{opt}$).

Parameters of MLR models were estimated with the least square method using Excel. The model accuracy was evaluated using the coefficient of determination $R^2$ and adjusted $R^2$. In 6 regression models, the statistical significance of regression equation and regression coefficients was assessed by F-test and student t-test, respectively and the difference of Pearson's correlation coefficients in 6 regression models was tested based on Fisher's transformation [16].

### 2.2.2. *Construction of ANN models for the prediction of SBE efficiencies of single MM primers and training*

ANN models corresponding six MLR models were constructed to improve the prediction accuracy of MLR models.

As shown in Fig. 1, ANNs were composed of an input layer, a hidden layer and an output layer. The hidden layer and the output layer contained four neurons (considering the size of data) and a neuron, respectively. The input layers of ANN models were constructed differently each other according to MLR models described before. For example, the input layer of ANN model 1 contained 19 neurons representing the position effect of MMs in primer-template duplex as regression model 1 and the input layer of ANN model 2 contained 19 neurons of ANN model 1 and a neuron representing the annealing temperature as regression model 2. And input layers of ANN model 5 and ANN model 6 contained 31 and 32 neurons equal to numbers of independent variables of regression model 5 and 6, respectively.

Data prepared for 4 MLR models with the high prediction accuracy were used for training, validation and testing of corresponding ANN models. Of 227 data, 159 input-output data (70%) were used for training and 34 input-output data (15%) selected randomly were used for validation and testing, respectively.

Neural Network Fitting (nntraintool) of Matlab 2018 was selected as ANN tool and was trained by Bayesian Regularization (trainbr). As shown in Fig. 2, during training of the ANN model, the weights and biases (also known as the adjustable network parameters) are adapted until the network output matches the target [15]. This procedure is undertaken to enhance the network performance. The training step can be completed if the magnitude of the gradient of performance is lower than 1e −3 or epoch reaches to 1,000 [15]. ANN model that shows the highest Pearson correlation coefficient R between Outputs and Targets in all data set was selected [15]. The significance of the difference of Pearson correlation coefficients in four different ANN models were tested based on Fisher's transformation [16].

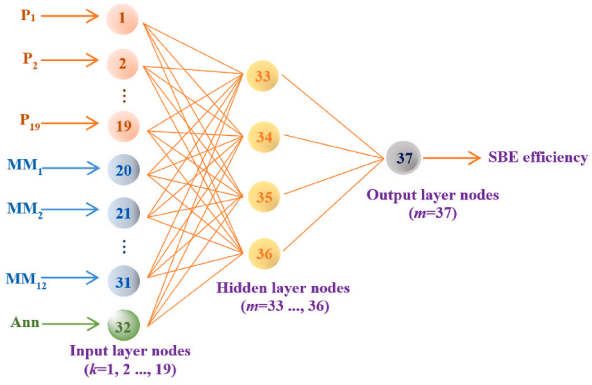



**Fig. 1.** ANN(for example, ANN model 6) with a structure of 32–4–1 for the prediction of SBE efficiencies of MM primers.

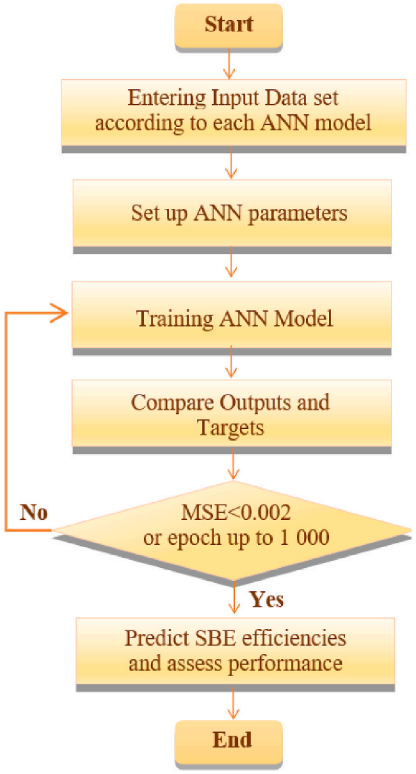



**Fig. 2.** Predication flowchart of SBE efficiencies of MM primers by ANN MSE: mean squared error.

## 3. Result

### 3.1. MLR models for the prediction of SBE efficiencies of single MM primers

First of all, 6 regression models with the independent variables representing the position effect of MMs, type of MMs or annealing temperature and a dependent variable representing the SBE efficiencies of primers were suggested and the regression analysis was conducted in order to establish the linear regression models predicting experimental data that shows the SBE efficiencies of primers according to primer-template MMs and annealing temperature in two primers reported by Wu et al. [10] (Table 1).

As Table 1 shows, regression model 1 considering only the position effect of MM in primer-template duplexes (formula 1) showed the relatively high coefficient of determination ($R^2 = 0.5466$) in despite of linear regression model, and regression model 2 ($R^2 = 0.6731$) considering both the position effect of MMs and annealing temperature (formula 2) showed the highly significant $R^2$, compared to regression model 1. This means that the position effect of MM plays the most important role in predicting of SBE efficiencies of primers and the annealing temperature is a factor to improve significantly the predictability of linear regression models. Regression model 3 considering only the type of MMs in primer-template duplexes (formula 3) showed the worst result among regression models tested, though the regression equation ($p = 4.245E\text{-}06 < 0.05$) and the correlation coefficient ($R = 0.4406$, $p = 1.4593E\text{-}12 < 0.001$) were significant, respectively. The $R^2$ of regression model 4 considering the type of MMs and annealing temperature (formula 4) and regression model 5 considering the position of MMs and the type of MMs (formula 5) were 0.3286 and 0.6578

**Table 1**
Results of MLR analysis and significance of the differences between regression coefficients in the 6 regression models.

| Regression Models | | Linear Regression Analysis | | | | | | R※ |
|---|---|---|---|---|---|---|---|---|
| No | Variables | $R^2$ | Adjusted $R^2$ | $SS_{resid}$ | *f* | F | *p* value | |
| 1 | 19 | 0.5466 | 0.5050 | 5.9540 | 207 | 13.1331 | 2.906E-26 | $0.7393^{c}$ |
| 2 | 20 | 0.6731 | 0.6414 | 4.2923 | 206 | 21.2106 | 1.558E-39 | $0.8204^{b}$ |
| 3 | 12 | 0.1941 | 0.1489 | 10.583 | 214 | 4.2952 | 4.245E-06 | $0.4406^{d}$ |
| 4 | 13 | 0.3286 | 0.2877 | 8.8158 | 213 | 8.0205 | 5.276E-13 | $0.5733^{d}$ |
| 5 | 31 | 0.6578 | 0.6034 | 4.4939 | 195 | 12.0901 | 6.092E-31 | $0.8110^{b,c}$ |
| 6 | 32 | 0.7757 | 0.7388 | 2.9448 | 194 | 20.9717 | 3.827E-47 | $0.8808^{a}$ |

※R is correlation coefficient between Outputs and Targets, and different letters such as a, b, c and d indicate that Rs have significant differences ($p < 0.05$). ※$SS_{resid}$ is the residual sum of squares.

respectively, and they have no significant differences compared to regression model 3 ($p = 0.0576$) and regression model 1 ($p = 0.0553$). But regression model 6 ($R^2 = 0.7757$) considering all 3 parameters (position of MMs, type of MMs and annealing temperature) (formula 6) showed the best result among regression models tested, and also showed significant difference, compared to regression model 2 ($p = 0.0193 < 0.05$). This means that the type of MMs itself has no big effect but it exerts a positive influence upon the predictability of the regression models when it cooperates with the position of MMs and annealing temperature.

Regression lines reflecting the relation between the relative efficiencies of SBE (Target) and predictions of regression models (Output) obviously shows the fact that regression model 6 considering all 3 parameters reproduces the SBE efficiencies of MM primers more correctly than other regression models considering 1 or 2 parameters (Fig. 3). The regression model 2 considering only 2 parameters (position of MMs and annealing temperature) reproduces the best SBE efficiencies of primers next to the regression model 6.

The regression equation of the model 6 obtained by the least squares method is denoted as follows:

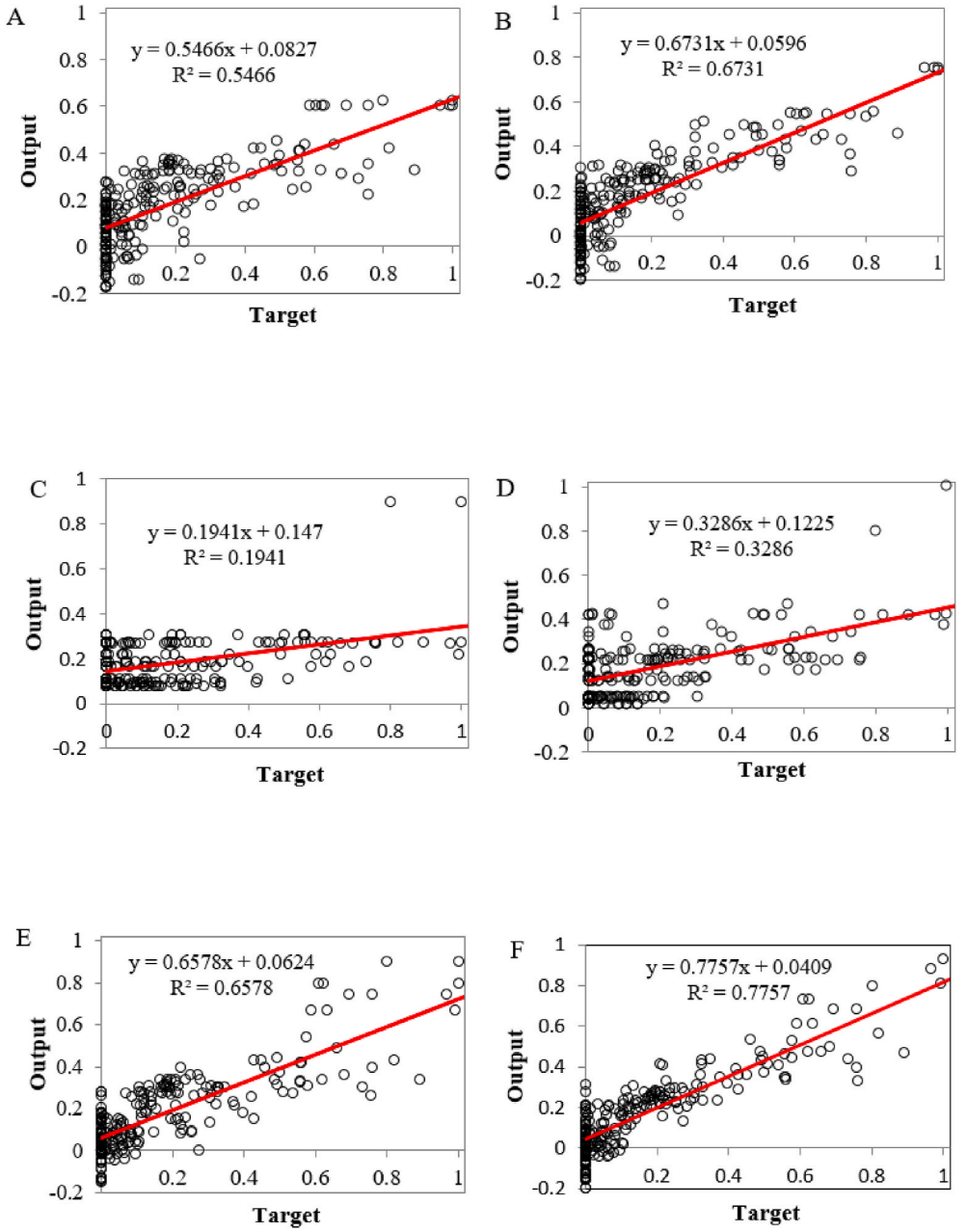



**Fig. 3.** Correlation between the relative extension efficiencies of primers and the predictions in various regression models. (A) regression model 1, (B) regression model 2, (C) regression model 3, (D) regression model 4, (E) regression model 5, (F) regression model 6.

$$y = -12.334 + 1.980P_1 + 0.893P_2 + 0.736P_3 + 0.848P_4 + 0.796P_5 + 0.613P_6 + 0.711P_7 + 0.678P_8 + 0.561P_9 + 0.820P_{lO} + 0.726P_{ll} + 0.623P_{l2} + 0.589P_{l3} + 0.260P_{l4} + 0.449P_{l5} + 0.438P_{l6} + 0.193P_{l7} + 0.409P_{l8} - 0.012P_{l9} - 0.357MM_{l(A:A)} - 0.292MM_{2(A:C)} - 0.284MM_{3(A:G)} - 0.122MM_{4(C:A)} - 0.073MM_{5(C:C)} - 0.193MM_{6(C:T)} - 0.366MM_{7(G:A)} - 0.395MM_{8(G:G)} - 0.453MM_{9(G:T)} - 0.301MM_{lO(T:C)} - 0.295MM_{ll(T:G)} - 0.410MM_{l2(T:T)} + 1.025Ann. \quad (7)$$

As formula 7 shows, regression coefficients $b_1$ ($b_{P1}$) – $b_{18}$ ($b_{P18}$) representing the position effect of MMs were positive in regression model 6: regression coefficient $b_1$ ($b_{P1}$) representing the position effect of 3′-terminal MMs has the biggest value among them and others tend to being reduce going to 5′-end with some fluctuation. This reproduces well the fact [10] that MMs nearby 3′-end of primer exert the most remarkable influence upon SBE efficiencies of MM primers and the effect is reduced gradually going to 5′-end of primer. Furthermore, all regression coefficients $b_{20}$ ($b_{MM1(A:A)}$) –$b_{31}$ ($b_{MM12(T:T)}$) representing the type of MMs were negative in regression model 6: this also reproduces well the fact that all types of MMs depress the stability of primer-template duplex and then reduces the SBE efficiencies of primers than PMs, comparing the quantitative differences of Gibbs free energy changes corresponding to the nearest neighbor MMs reported by SantaLucia [7]. Regression coefficient $b_{32}$ ($b_{Ann}$) representing the effect of annealing temperature is a large positive value and it reproduces the fact [10] that annealing temperature affects the SBE efficiencies of primers considerably and the relative SBE efficiency has the maximum value at optimal annealing temperature. Like this, it could be considered that regression coefficients of regression model 6 obviously illustrate the effect of position and type of MMs and the annealing temperature on SBE efficiencies in primer-template duplex.

In regression equation (7), all regression coefficients were significant ($p < 0.05$) except a regression coefficient $b_{19}$ ($b_{P19} = -0.0117$, $p = 0.5958 > 0.05$) corresponding to 19th position from 3′-end of primer and 3 regression coefficients $b_{23}$ ($b_{MM4(C:A)} = -0.1222$, $p = 0.2200 > 0.05$), $b_{24}$ ($b_{MM5(C:C)} = -0.0730$, $p = 0.4674 > 0.05$) and $b_{25}$ ($b_{MM6(C:T)} = -0.1934$, $p = 0.0512 > 0.05$) corresponding to three types of MMs.

In Table 1, adjusted $R^2$ of regression model 2 considering both the annealing temperature and the position effect of MMs is larger than that of regression model 1 considering only the position effect of MMs but smaller than that of model 6 which considered additionally 12 types of MMs. The significant differences between regression model 1 and regression model 2 ($p = 0.0268 < 0.05$) and between regression model 2 and regression model 6 ($p = 0.0193 < 0.05$) were confirmed by the test based on Fisher's transformation. This means that regression model 6 is the best regression model for prediction of SBE efficiencies of MM primers.

Despite of linear regression model, determination coefficients of regression model 2, 5 and 6 are all ($R^2 > 0.65$) significantly higher than that of polynomial regression model reported by Wu et al. [10] ($R^2 = 0.5318$) showing the relation between the relative extension efficiency of primers and the melting temperature difference of PM and MM duplex ($\Delta T_m$).

In conclusion the regression model considering the effects of position and type of MMs, and annealing temperature reproduces the relative extension efficiencies of MM primers well and its prediction accuracy is higher than the polynomial regression model suggested by Wu et al. [10].

### 3.2. ANN models for prediction of SBE efficiencies of single MM primers

MLR models for prediction of relative extension efficiencies of single MM primers are better than polynomial regression model reported by Wu et al. [10]. But its prediction accuracy is not enough because the relations between independent and dependent variables in the linear regression models are not just linear in reality. In order to improve the prediction accuracy of models, we selected 4 regression models in 6 regression models. Among 6 regression models, model 3 and model 4 have been excluded because of their low $R^2$. Independent and dependent variables of regression models were used as input and output data in corresponding ANN models and the ANN models were trained with Bayesian regularization method (Table 2).

In 4 ANN models, the correlation coefficient between relative extension efficiencies of primers (Target) and predictions of models (Output) were all obviously improved in comparison with corresponding regression models (Tables 1 and 2).

The test result based on Fisher's transformation shows the significant difference ($p = 2.22E\text{-}19 < 0.001$) between correlation coefficients of ANN model 1 (using 19 neurons representing the position effect of MMs in input layer) and ANN model 2 (using 20 neurons representing the position effect of MMs and the annealing temperature in input layer). The significant difference also exists between ANN model 5 and ANN model 6 ($p = 5.070E\text{-}07 < 0.001$). The correlation coefficient of ANN model 5 using 31 neurons representing the effects of position and types of MMs was significantly lower than that of ANN model 2 ($p = 8.182E\text{-}13 < 0.001$) and had no significant difference in comparison with that of ANN model 1 ($p = 0.3007 > 0.05$). In ANN model 6 showing the highest prediction accuracy of the relative extension efficiencies of MM primers, the correlation coefficient for training set, testing set and all

**Table 2**
Training results in various ANN models and statistical significance of differences between the correlation coefficients of ANN models.

| ANN Models | Input Layer | Hidden Layer | Training | | Testing | | All |
|---|---|---|---|---|---|---|---|
| | | | R | MSE | R | MSE | R※ |
| 1 | 19 | 4 | 0.8419 | 0.0178 | 0.8082 | 0.0136 | $0.8386^c$ |
| 2 | 20 | 4 | 0.9622 | 0.0045 | 0.9726 | 0.0027 | $0.9634^b$ |
| 5 | 31 | 4 | 0.8856 | 0.0131 | 0.7666 | 0.0226 | $0.8654^c$ |
| 6 | 32 | 4 | 0.9870 | 0.0016 | 0.9782 | 0.0022 | $0.9857^a$ |

※ R is correlation coefficient between Outputs and Targets, and different letters such as a, b and c indicate that Rs have significant differences ($p < 0.05$).

data set were 0.9870($R^2 = 0.9742$), 0.9782($R^2 = 0.9569$) and 0.9857 ($R^2 = 0.9716$), respectively. These values are clearly higher than those of other ANNs ($p$-value$<0.001$).

As shown in Fig. 4, regression lines representing the relationship between relative extension efficiencies (Target) and predictions of models (Output) indicate that ANN model 6 (considering the effect of position and type of MMs and annealing temperature) is the best model, and that ANN model 2 (considering the position effect of MMs and annealing temperature) is better than ANN model 1 or ANN model 5. Furthermore, the prediction accuracy of SBE efficiencies of primers of ANN model 6 was considerably higher than that of the regression model 6.

As shown in Fig. 2, there are 2 outliers on the Target-Output plot of ANN model 6. These outliers are responsible for 54th and 168th input data that contain the same MM(C:C) at the same position (the 18th position from primer 3′end). Here, two primers are the completely identical MM primers (5′-GCCTACCTTGTTACGACTT-3′) that designed from primer 1492R (5′-GGCTACCTTGTTACGA CTT-3′) and the annealing temperatures are all 60 °C. But only the templates are different each other; 16S-ITS-23S gene fragment of *Bacteroides coagulans* in 54th input and 16S rRNA fragment of *Stenotrophomonas acidaminiphila* in 168th input and the relative SBE efficiencies in two input data were 0.7591 and 0.2236, respectively and have the significant difference. The predicted efficiencies responsible for 54th and 168th inputs were 0.3999 in ANN model 6 and were shown as 2 outliers in Fig. 2. It seems that the MM primers with the instable MMs such as C:C are seriously affected by other factors like the sequences and structures of different templates.

Thus, ANN model 6 is promising in the design and the specificity test of primers as it reflects the quantitative relationship between MMs of primer-template duplex and relative extension efficiency of primers and its prediction accuracy is much higher than the polynomial equation reported by Wu et al. [10].

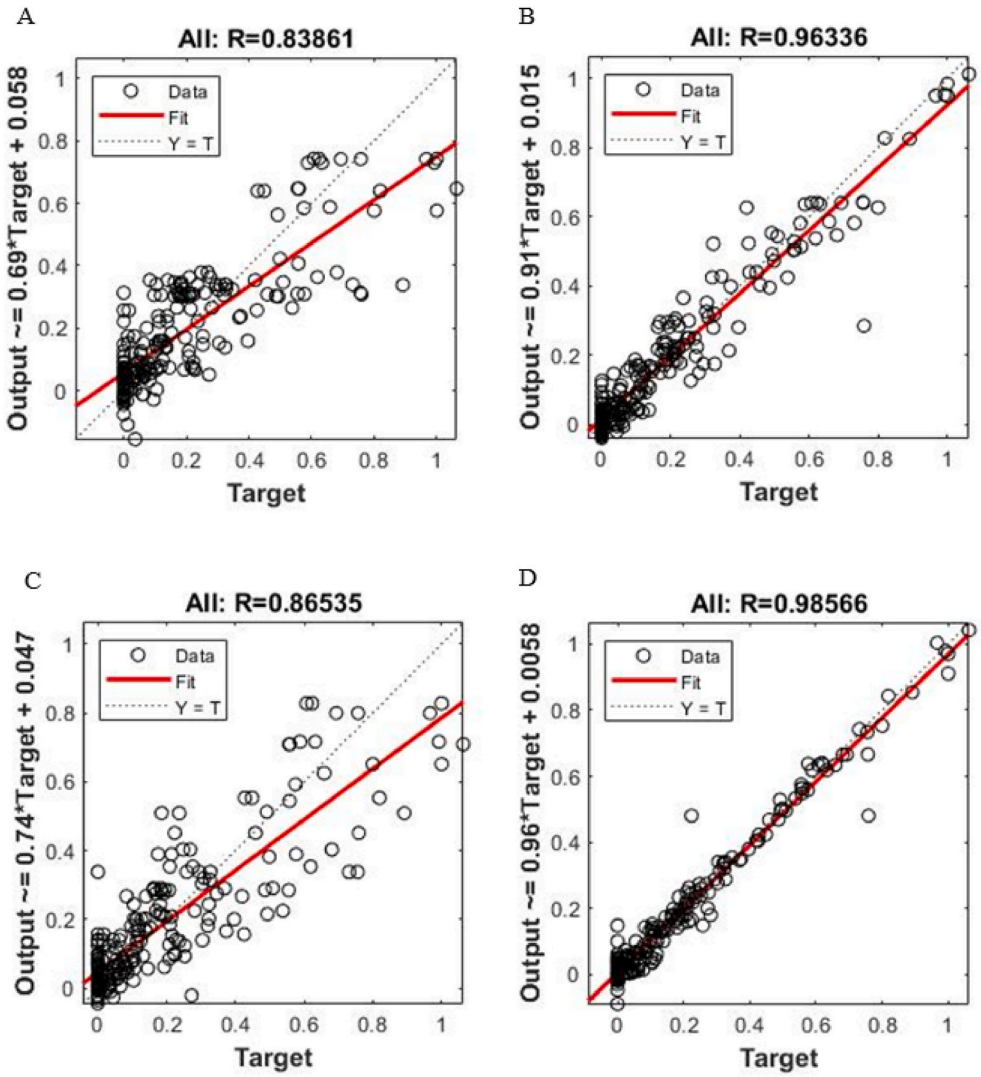



**Fig. 4.** Correlations between the relative extension efficiencies of primers and the predictions in various ANN models. (A) ANN model 1, (B) ANN model 2, (C) ANN model 5, (D) ANN model 6.

## 4. Discussion

The extension efficiencies of primers are determined by the stability of primer-template duplex and the selective binding between DNA polymerase and 3′-end of primer-template duplex. MMs in primer-template duplex can affect the selective binding between DNA polymerase and primer-template duplex as well as the stability of duplex, and they can further affect the extension efficiencies of primers.

Wu et al. [10] determined the SBE efficiencies of two sets of primers (PM primers and possible primers including all types of single MMs) using the SBE assay and illuminated the effect of the position and the types of primer-template MMs on SBE efficiencies. In general, the relative extension efficiencies of MM primer-template duplexes were increased as the position of MMs moves to 5′ end. As the position of MMs moves from the 5th position to 5′end, it was observed that the MMs with low $\Delta T_m$ usually result in the high extension efficiency. And the SBE efficiencies of primers with the MMs such as G:G, T:G, G:T, A:G and G:A, that make the primer-template duplex relatively stable, were higher than that of primers with the MMs such as T:T, A:A, C:T, T:C, C:A, A:C and C:C, that make the duplex relatively instable. When MMs such as G:A, G:T locates at the 7th, 8th, 9th, 12th, 15th,16th position of primer 1492R and at the 10th, 15th, 16th position of primer MX, primers have relatively high SBE efficiencies. These experimental data show that the negative correlation between $\Delta T_m$ ($Tm$ difference between PM and MM duplexes) and SBE efficiencies of primers and the more improved cubic polynomial equation with determination coefficient $R^2 = 0.5318$ ($R = -0.7292$) could be obtained by using the polynomial regression equation.

We tested MLR models and ANN models predicting the SBE efficiencies of primers, based on the assumption that SBE efficiencies of primers are associated with not only $\Delta T_m$, the difference of melting temperature between PM and MM primer-template duplex, but also the position and type of MM and annealing temperatures. It was shown that both the MLR model 6 and the ANN model 6 reproduce SBE efficiencies of primers with higher accuracy in comparison with polynomial regression model showing the relation between $\Delta T_m$ and SBE efficiencies of primers.

This means that independent variables of MLR model 6 and input data of ANN model 6 were selected and encoded well to reproduce the relation between MMs of primers and SBE efficiency.

The coefficient of determination $R^2$ cannot be reduced as the independent variables are added to model, and the model that gives the maximum $R^2$ is just the model that contains all independent variables. But the model with maximal $R^2$ is not always the best one. The adjusted $R^2$ does not always increase as variables are added to model because the impact of degrees of freedom was removed. So the adjusted $R^2$ is more suitable than $R^2$ for evaluation of models containing different number of independent variables. In other words, the model that gives the maximum adjusted $R^2$ can be chosen as the "best" model [16].

In regression models, model 6 is the "best model", for it has the biggest adjusted $R^2$ and the differences of correlation coefficients between the relative extension efficiencies of primers and predictions of models are significant.

In order to express the position effect of MMs in primer-template duplexes, each position has to be represented by 4 signs corresponding to 4 nucleotides. 4 neurons can be assigned on each position of duplexes in ANN models and denoted by either 1 or 0 according to the presence or absence of MMs in each position of duplexes, or by position-specific score matrix (PSSM) representing the frequency of existence of given nucleotides in multiple alignments. However, these methods need much more training data due to its large number of neurons, and variables or neurons representing the position effect of MMs may have no relations with the stability of primer-template duplexes. We tried to denote quantitatively the extent that MMs at given position contribute to the stability of primer-template duplexes by normalizing the differences of Gibbs free energy changes between two flanking MM doublets (nearest neighbors) in the primer-template duplex and their corresponding PM doublets, and tried to express the position effect of MMs with the minimum number of independent variables or neurons. In primer-template duplexes, the difference of Gibbs free energy changes between the nearest neighbor MM doublet and its corresponding PM doublet represents the hydrogen bond and dimer-stacking of MMs, and the sequence contexts of primer-template duplexes are also considered to some extent since it uses two nearest neighbors centered to the position of MMs.

Gibbs free energy changes of nearest neighbor MMs have been widely used for predicting the melting temperature of primers [4,5,8–10], but never have been used for predicting the SBE efficiencies of MM primers by encoding the position effect of MMs. Therefore, this method is a way that reduces the number of independent variables in regression models or the number of neurons of input layer in ANN models considering the position effect of MMs on duplex stability and the doublet contexts of primer-template duplex, and it can improve the prediction accuracy of models.

The type of MMs itself cannot tell the SBE efficiencies of primers accurately, but it can significantly improve the prediction accuracy of model when combined with the position effect of MMs and the annealing temperature. The encoding by the normal distribution centered at the optimal annealing temperature is also the reliable way that can apply to the amplification of primer extension reaction by considering the changes of SBE efficiencies of primers according to the annealing temperature.

ANN model 6 can predict SBE efficiencies of primers with even higher accuracy ($R > 0.98$, correlation coefficient between relative extension efficiencies and predictions of models). This means that there exists a non-linear relation between the SBE efficiencies of primers and the parameters such as the position effect and the type of MMs in primer-template duplex, and the annealing temperature.

The prediction model proposed in this paper, which considered the position effect and the type of MM and the annealing temperature, can predict the primer extension efficiencies with high accuracy when the sequence of the primer and the annealing temperature are given. So we think that this model can be widely used for the design of efficient primers and for the specificity test of primers in genome databases.

We are sure that even more comprehensive prediction models representing the effect of multiple MMs on PCR efficiencies will be obtained if related experimental data are accumulated in the future.

## 5. Conclusions

MLR models and ANN models for the prediction of SBE efficiencies of primers with a single MM were developed. It is apparent that 19 variables representing the position effect of MMs, 12 variables representing the type of MMs and a variable representing annealing temperatures are important and valuable for predicting SBE efficiencies of MM primers from primer-template sequences and annealing temperatures in MLR model 6. ANN model 6 with 32 neurons representing the position effect and the type of MMs in primer-template duplexes and the annealing temperature in input layer showed the highest prediction accuracy (correlation coefficients R were 0.9870 ($R^2 = 0.9742$), 0.9782($R^2 = 0.9569$) and 0.9857 ($R^2 = 0.9716$), respectively in training set, testing set and all data sets). This model can predict the SBE efficiencies of primers from the primer sequence and the annealing temperature with high accuracy.

## Author contribution statement

Conceived and designed the experiments, Wrote the paper: Myong-Rim Ri.
Performed the experiments: Jin-Sok Kang.
Analyzed and interpreted the data:, Myong-Ryong Ri, Contributed reagents, materials, analysis tools or data: Song Nam U.
All authors have read and agreed to the final version of the manuscript.

## Data availability statement

No data was used for the research described in the article.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] D. Bru, et al., Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example, Applied and Environmental Microbiology, Mar. 1660–1663 (2008), https://doi.org/10.1128/AEM.02403-07.
[2] L.M. Oliveira, et al., Melting temperature measurement and mesoscopic evaluation of single, double and triple DNA mismatches, Chem. Sci. 11 (2020) 8273–8287, https://doi.org/10.1039/D0SC01700K.
[3] P. Miranda, G. Weber, Thermodynamic evaluation of the impact of DNA mismatches in PCR-type SARS-CoV-2 primers and probes, Mol. Cell. Probes 56 (4) (2021), 101707, https://doi.org/10.1016/j.mcp.2021.101707.
[4] J.C.O. Guerra, P. Licinio, Terminal contributions for duplex oligonucleotide thermodynamic properties in the context of nearest-neighbor models, Biopolymers 95 (3) (2011) 194–201, https://doi.org/10.1002/bip.21560.
[5] C.A. Johnson, et al., Computational model for predicting experimental RNA and DNA nearest-neighbor free energy rankings, J. Phys. Chem. B 115 (2011) 9244–9251, https://doi.org/10.1021/jp2012733.
[6] J. SantaLucia Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. USA 95 (1998) 1460–1465, https://doi.org/10.1073/pnas.95.4.1460.
[7] J. SantaLucia Jr., D. Hicks, The thermodynamics of DNA structural motifs, Annu. Rev. Biophys. Biomol. Struct. 33 (2004) 415–440, https://doi.org/10.1146/annurev.biophys.32.110601.141800.
[8] C.B. Hughesman, et al., A new general model for predicting melting thermodynamics of complementary and mismatched B - form duplexes containing locked nucleic acids: application to probe design for digital PCR detection of somatic mutations, Biochemistry 54 (2015) 1338–1352, https://doi.org/10.1021/bi500905b.
[9] C.B. Hughesman, et al., Role of the heat capacity change in understanding and modeling melting thermodynamics of complementary duplexes containing standard and nucleobase-modified LNA, Biochemistry 50 (2011) 5354–5368, https://doi.org/10.1021/bi200223s.
[10] J.H. Wu, et al., Quantitative effects of position and type of single mismatch on single base primer extension, J. Microbiol. Methods 77 (2009) 267–275, https://doi.org/10.1016/j.mimet.2009.03.001.
[11] S. Kwok, et al., Effects of primer - template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies, Nucleic Acids Res. 18 (4) (1990) 999–1005, https://doi.org/10.1093/nar/18.4.999.
[12] B.M. Ledeker, S.K. De Long, The effect of multiple primer–template mismatches on quantitative PCR accuracy and development of a multi-primer set assay for accurate quantification of pcrA gene sequence variants, J. Microbiol. Methods 94 (2013) 224–231, https://doi.org/10.1016/j.mimet.2013.06.013.
[13] R. Stadhouders, 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) (2010) 109–117, https://doi.org/10.2353/jmoldx.2010.090035.
[14] S. Kouadri, et al., Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models, Environ. Sci. Pollut. Control Ser. 29 (2022) 21067–21091, https://doi.org/10.1007/s11356-021-17084-3.
[15] A. Hamdy, et al., Regression analysis and artificial intelligence for removal of methylene blue from aqueous solutions using nanoscale zero-valent iron, Int. J. Environ. Sci. Technol. 16 (2019) 357–372, https://doi.org/10.1007/s13762-018-1677-z.
[16] B. Vidakovic, Chapter 15 correlation, in: Statistics for Bioengineering Sciences with MATLAB and WinBUGS Support", Springer, 2011, pp. 571–597, https://doi.org/10.1007/978-1-4614-0394-4_15.