Hindawi
Journal of Healthcare Engineering
Volume 2022, Article ID 8739447, 5 pages
https://doi.org/10.1155/2022/8739447

Research Article

Monitoring and Early Warning Analysis of the Epidemic Situation of *Escherichia coli* Based on Big Data Technology and Cloud Computing

Meishu Yan[1] **and Meizi Yan**[2]

[1]*Affiliated Hospital of Chengde Medical College, Chengde, Hebei 06700, China*
[2]*Chengde Medical College, Chengde, Hebei 067000, China*

Correspondence should be addressed to Meizi Yan; ymz_2016@cdmc.edu.cn

Received 6 November 2021; Accepted 21 January 2022; Published 9 February 2022

Academic Editor: Kalidoss Rajakani

The purpose of this study is to analyze the molecular epidemiological characteristics and resistance mechanisms of *Escherichia coli*. The study established a big data cloud computing prediction model for the epidemic mechanism of the pathogen. The study establishes the early warning, control parameters, and mathematical model of *Escherichia coli* infectious disease and monitors the molecular sequence of the pathogen based on discrete indicators. A nonlinear mathematical model equation was used to establish the epidemic trend model of *Escherichia coli*. The study shows that the use of the model can control the relative error at about 5%. The experiment proves the effectiveness of the combined model.

1. Introduction

Gene promoter is the most important regulatory element of gene transcription; it determines where the gene expression starts. Therefore, the study of promoters has always been a hot spot in modern molecular biology. The theoretical prediction of gene promoters has become an important research content of bioinformatics as an important part of the identification of the complete structure of genes [1]. With the advent of the postgenomic era, although a large amount of genomic data have been generated, the available annotation information related to the promoter is still relatively scarce. Therefore, it is urgent to design a fast and effective method to identify the promoter sequence in the genome.

Because prokaryotes and eukaryotes genome promoters are quite different, they are usually studied separately for prediction. *Escherichia coli* is one of the most important prokaryotic model organisms. At present, a variety of mathematical models have been used to predict the promoter of *Escherichia coli*. The position weight matrix (PWM) is a more commonly used prediction method. Some scholars selected 288 different PWMs to conduct a systematic study on 599 sigma70 promoters. The study found that the sensitivity reached 86%, while the accuracy rate was only 53%. Some scholars have predicted 469 *Escherichia coli* promoter sequences and their positions based on predicted transcription units and using the Markov model (MM). The accuracy rate is more than 70%. The neural network method (NN) has also been used many times to predict the promoter of *Escherichia coli*. Recently, some scholars have used NNPP2.2 software to combine the distance from TSS to the translation initiation site (TIS) to improve the prediction accuracy of the *Escherichia coli* promoter [2]. Some scholars used the support vector machine (SVM) to predict 669 *Escherichia coli* sigma70 promoters and obtained high prediction accuracy. Some scholars have proposed a prokaryotic promoter identification method based on feature screening, and this method has also achieved satisfactory prediction results. We once proposed a position association scoring matrix (PCSM) algorithm to improve the prediction accuracy of promoters. Recently, some scholars have

obtained higher recognition accuracy by combining the diversity increment with the secondary discriminant analysis (IDQD) method.

Although the prediction success rate of promoters is constantly improving, there are still many problems. First of all, the promoter datasets used in the past are mostly small, and the nonpromoter datasets are relatively large. This will increase the number of false positives and affect the accuracy of performance evaluation. Second, most of the work does not have a deep understanding of promoters and insufficient utilization of characteristic information [3]. Again, most of the work has carried out two predictions such as promoter and gene and promoter and coding region, and the actual need is to identify the promoter sequence from the entire genome. Therefore, such predictions lack practical significance.

In view of the problems in the prediction of the above promoters, this article will reintegrate and predict the characteristics of the promoter sequence of *Escherichia coli*. First, consider the interaction between RNA polymerase and promoter sequence. We use the position association scoring function (PCSF) to describe the positional conservation of promoter sequences. Second, the promoter sequence is divided into different windows according to the sequence characteristics, and the discrete increment index (ID) is used to measure the information content of the sequence in each window. Finally, we used the modified Markov discriminant to predict the promoter of *Escherichia coli*. Here, we call this method the IPMD algorithm [4]. Comparison with previous results shows that the algorithm we developed has better predictive performance and is more practical.

2. Materials and Algorithms

2.1. The Establishment of the Database. The *Escherichia coli* sigma70 promoter sequence is from Regulon DB, an annotation database of the *Escherichia coli* transcription regulation network. A total of 741 experimentally confirmed sigma70 promoters were obtained, and the length of each promoter sequence was 81 bp (−60...+20, TSS reference is 0 position). The negative dataset was obtained from the whole genome of *Escherichia coli* (downloaded from GenBank, sequence AC number U00096) without the promoter. But in fact, there is no experiment to prove which part of the sequence does not contain a promoter [5]. Therefore, according to the known transcription unit structure of *Escherichia coli* and the known promoter or coding region location, try to avoid regions where promoters may appear to extract negative data. The nonpromoter sequence selected in this study comes from two regions: coding region sequence and noncoding region sequence. Since the promoter drives its downstream genes, it is generally located at the head of the coding region. However, because the *Escherichia coli* genome is small, 89% are coding regions, so some promoters will exist at the end of the previous gene. Therefore, the nonpromoter of the coding region is selected in the middle part of the longer gene. Next, we select nonpromoter sequences from noncoding regions.

Based on the above considerations, we selected 700 nonpromoter and 700 nonpromoter sequences in the coding region and 700 nonpromoter sequences in the convergent region, each of which was 81 bp in length.

2.2. Location-Related Scoring Function. Define the standard sample set as $\sum$ and the position correlation weight matrix as $P = [p_{xi}]_{M\times L}$, where M is the number of types of characters, L is the length of the sequence, and p_{xi} represents the probability of character x appearing at position i. $p_{xi} = n_{xi}/N_i$, N_i is the number of the sequence, $N_i = \sum_x n_{xi}$.

Count the number of sixet fragments at each position in the sequence. We introduce the pseudocount B_i and redefine the matrix elements of the position association weight matrix as

$$p_{xi} = \frac{\left(n_{xi} + p_0\sqrt{B_i}\right)}{\left(N_i + B_i\right)}, \tag{1}$$

where p_0 is the background frequency, defined as $P_0 = 1/N_i$. We use the position weight matrix, and the associated scoring function is defined as

$$F = \sum_i \ln\left(\frac{P_{xi}}{P_0}\right). \tag{2}$$

The value of F is used to characterize the degree of similarity between a sequence and a promoter sequence [6]. The larger the value of F, the more likely this sequence is a promoter sequence.

2.3. Discrete Increment. If there are two datasets X: $[n_1, n_2, \ldots, n_s]$, Y: $[m_1, m_2, \ldots, m_s]$, the discrete increment is defined as

$$\begin{aligned}
&\Delta(X,Y) = D(X+Y) - D(X) - D(Y) = D(N,M) - \sum_{i=1}^{s} D\left(n_i, m_i\right),\\
&N = \sum_{i=1}^{s} n_i,\\
&M = \sum_{i=1}^{s} m_i,\\
&D(N,M) = (N+M)\log_b(N+M) - N\log_b N - M\log_b M,\\
&D\left(n_i, m_i\right) = \left(n_i + m_i\right)\log_b\left(n_i + m_i\right) - n_i\log_b n - m_i\log_b m_i.
\end{aligned} \tag{3}$$

If n_i or m_i is zero, then $D(n_i, m_i) = 0$. It is easy to prove that the discrete increment is nonnegative, namely, $\Delta(X,Y) \geq 0$. We take the natural logarithm (in this case, the unit of information is knight). The discrete increment $\Delta(X,Y)$ can be regarded as a quantitative expression of the biological similarity relationship, which reflects the similarity of the two sets of data [7]. The smaller the $\Delta(X,Y)$, the more similar the two sets of data.

2.4. Modified Markov Discriminant. Considering samples with multiscale features, this study uses modified Markov discriminant to integrate features [8]. For any promoter

sequence S to be predicted, the discriminant function between it and the training set can be defined as

$$\mathrm{MD}(s,\mu) = (s-\mu)^T C^{-1}(s-\mu) + \lg|C|. \quad (4)$$

Then, the type of sequence S can be given by the following discriminant rules:

$$\xi = \mathrm{MD}\left(s,\mu_{\text{pranoter}}\right) - \mathrm{Min}\left\{\mathrm{MD}\left(s,\mu_{\text{coding}}\right),\mathrm{MD}\left(s,\mu_{\text{nin-coding}}\right)\right\}. \quad (5)$$

Operator Min represents the smallest value in the brackets. The type of the sequence to be tested for any given threshold ξ_0 can be predicted.

2.5. Accuracy Evaluation. We use the definitions of sensitivity (S_n), specificity (S_p), and correlation coefficient (CC) to evaluate the predictive performance of the algorithm.

$$\begin{aligned}
\text{Senstivity: } S_n &= \frac{\mathrm{TP}}{\mathrm{AP}},\\
\text{Specificity: } S_p &= \frac{\mathrm{TP}}{\mathrm{AP}},\\
\text{False positive rate: } \mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{AN}},\\
\text{Total accuracy: } Ac &= \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})}.
\end{aligned} \quad (6)$$

The correlation coefficient $\mathrm{CC} = (\mathrm{TP}\circ\mathrm{TN} - \mathrm{FP}\circ\mathrm{FN})/\mathrm{PP}\circ\mathrm{PN}\circ\mathrm{AP}\circ\mathrm{AN}$; the abovementioned index is used to evaluate the standard of algorithm pros and cons.

Among them, $\mathrm{PP} = \mathrm{TP} + \mathrm{FP}$, $\mathrm{PN} = \mathrm{TN} + \mathrm{FN}$, $\mathrm{AP} = \mathrm{TP} + \mathrm{FN}$, $\mathrm{AN} = \mathrm{TN} + \mathrm{FP}$.

3. Forecast Results

3.1. Promoter Feature Selection. According to the sequence characteristics of the promoter of *Escherichia coli* and the conservative analysis of its promoter sequence in the past, the characteristics of the promoter of *Escherichia coli* were selected as follows:

(1) The conservative characteristic parameters of the promoter sequence. Select the sequence −51, −37, −36, −35, −34, −16, −15, −14, −13, −12, −11, −10, −9, −8, −7, −10, −2, −1 hexaplex of these 18 sites as the parameters of the positional association scoring function.

(2) The component characteristic parameters of the upstream promoter sequence. We select the frequency of hexaplexes between −60 bp and −25 bp in the sequence.

(3) Characteristic parameters of components near the transcription start site. We select the frequency of the hexat in the sequence between −25 bp and +21 bp.

Usually, the two-category problem has better prediction results than the three-category problem. However, because the negative data in the noncoding region and the negative data in the coding region are quite different in structure and composition, the two datasets are mixed into a negative dataset for promoter prediction research. This is bound to reduce the predictive performance of the model [9]. Therefore, the prediction model of this work will be generated by training on three datasets. The feature vector of the input modified Markov discriminant is a 9-dimensional vector (Table 1).

3.2. Forecast Accuracy. The prediction accuracy is the prediction ability of the test algorithm. We divide the positive sequence and the two types of negative sequence into two parts: the test set and the training set according to the ratios of 1 : 9, 2 : 8, 3 : 7, 4 : 6, and 5 : 5. In this way, the model is trained and tested [10]. The prediction results are given in Table 2. The results show that no matter what proportion of the IPMD model is trained and tested, its prediction accuracy has not changed significantly. This shows that our model is stable.

Although good prediction accuracy is obtained for various proportions of data, this test method does not fully reflect the predictive ability of the model. So next, we use a more objective 10-fold cross-check to evaluate the IPMD algorithm [11]. The 10-fold cross-check is to divide the dataset into 10 equal parts. We take one as the test set and the remaining 9 as the training set. This is repeated 10 times to test the algorithm. Then, use the receiver operating characteristic curve (ROC) to evaluate the algorithm performance. It is constructed by plotting the true positive rate and false positive rate calculated from a number of given thresholds. This is a comprehensive indicator that reflects the continuous changes in sensitivity and specificity. We use the area under the ROC curve to evaluate the prediction effect (Figure 1).

The results showed that the area under the ROC curve reached 0.953. When the optimal threshold ξ_0 is selected as −1.20, the prediction sensitivity reaches 84.9% and the specificity is 84.0%. The overall accuracy and correlation coefficient are 89.2% and 0.761, respectively.

3.3. Comparison of Results. The above only gives the prediction results of IPMD on the three datasets. Although the overall accuracy reaches about 90%, it is not certain that our model must be better than the prediction performance of other algorithms. Therefore, according to the previous prediction methods for promoters, we carried out prediction studies on the promoter and coding region sequence and the promoter and noncoding region sequence, respectively [12–14]. We compare this algorithm with other algorithms. The 10-fold cross-check is still used here, and the comparison results are given in Table 3. Our results have been further improved compared with previous algorithm results. This can prove that the prediction model that takes into account multiple characteristics can better identify the *Escherichia coli* sigma70 promoter [15].

Table 1: Promoter feature parameter selection.

Parameter	Source of information	PCSF or ID
PCSF promoter, PCSF coding, PCSF noncoding	Hexamer frequency in eighteen conservative sites	PCSF between test sequence and promoter, coding, noncoding set
ID1 promoter, ID1 coding, ID1 noncoding	Hexamer frequency in 60 bp: −25 bp	ID between test sequence and promoter, coding, noncoding set
ID2 promoter, ID2 coding, ID2 noncoding	Hexamer frequency in 25 bp: +20 bp	ID between test sequence and promoter, coding, noncoding set

Table 2: The influence of different ratios on the preresults of the IPMD model.

Ratio (test set : training set)	S_n (%)	S (%)	A_c (%)	CC
1 : 09	85	81	88	0.735
2 : 08	82	82	88	0.731
3 : 07	82	85	89	0.754
4 : 06	81	84	88	0.732
5 : 05	78	88	89	0.744

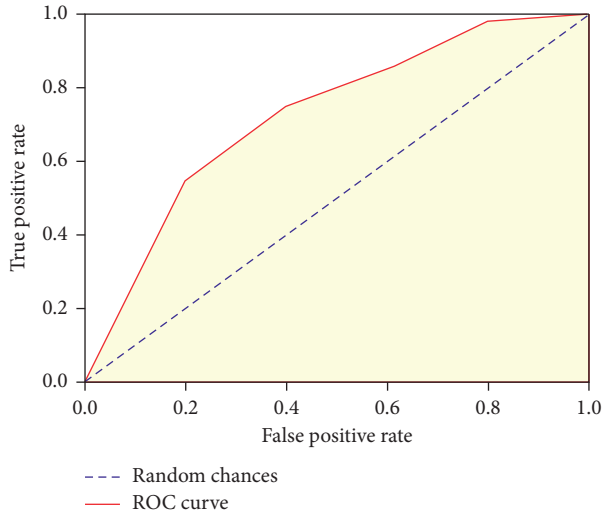

Figure 1: ROC curve predicted by the *Escherichia coli* sigma70 promoter.

Table 3: Comparison with the prediction results of other algorithms.

Method	S_n (×100%)	S_p (×100%)	CC
IPMD	95	91	0.844
	83	90	0.728
IDQD	94	83	0.75
	89	76	0.61
PCSM	91	81	0.68
	90	77	0.65
Sequence alignment kernel + SVM	82	84	0.67
	81	81	0.63
Boxes + SVM	76	83	0.62
	74	82	0.59
Boxes + threshold	76	83	0.61
	72	83	0.58
Zone likelihood + SVM	68	86	0.59
	67	84	0.56

4. Conclusion

In this study, a new prediction model of the *Escherichia coli* promoter is developed. We first considered the interaction between RNA polymerase and DNA sequence and constructed a position correlation scoring function. In fact, this scoring function can roughly measure the free energy of interaction between RNA polymerase and DNA sequence. Second, the discrete index is used to describe the sequence composition of different windows of the promoter. The discrete index is another reflection form of information entropy, so the discrete increment describes the increase of sequence information. Both have strict physics meaning, but they belong to different physics concepts, which can be regarded as orthogonal in mathematics. In this way, we got the promoter description method under multiple feature scales and then used the modified Markov discriminant to realize the promoter prediction of *Escherichia coli*. The comparison with other algorithms shows that our algorithm has better performance and stronger scalability and can be extended to the promoter prediction of other species.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Karkman, F. Berglund, C. F. Flach, E. Kristiansson, and D. G. J. Larsson, "Predicting clinical resistance prevalence using sewage metagenomic data," *Communications Biology*, vol. 3, no. 1, pp. 711–810, 2020.

[2] L. Aguirre, A. Vidal, C. Seminati et al., "Antimicrobial resistance profile and prevalence of extended-spectrum beta-lactamases (ESBL), AmpC beta-lactamases and colistin resistance (mcr) genes in *Escherichia coli* from swine between 1999 and 2018," *Porcine Health Management*, vol. 6, no. 1, pp. 8–6, 2020.

[3] A. Rahmani, L. Meradi, and N. C. E. H. Khennouchi, "Antimicrobial drug resistance of *Escherichia coli* strains isolated from urinary tract infection and raw beef meat in Oum El Bouaghi city of Algeria," *South Asian Journal of Experimental Biology*, vol. 11, no. 2, pp. 145–153, 2021.

[4] A. Indrawati, K. Khoirani, S. Setiyaningsih, U. Affif, S. Safika, and S. G. Ningrum, "Detection of tetracycline resistance genes among *Escherichia coli* isolated from layer and broiler

breeders in west java, Indonesia," *Tropical Animal Science Journal*, vol. 44, no. 3, pp. 267–272, 2021.

[5] A. F. A. Pires, J. Stover, E. Kukielka et al., "Salmonella and *Escherichia coli* prevalence in meat and produce sold at farmers' markets in northern California," *Journal of Food protection*, vol. 83, no. 11, pp. 1934–1940, 2020.

[6] S. P. Myoda, S. Gilbreth, D. Akins-Lewenthal, S. K. Davidson, and M. Samadpour, "Occurrence and levels of Salmonella, enterohemorrhagic *Escherichia coli*, and Listeria in raw wheat," *Journal of Food protection*, vol. 82, no. 6, pp. 1022–1027, 2019.

[7] S. Robatjazi, F. Nikkhahi, M. Niazadeh et al., "Phenotypic identification and genotypic characterization of plasmid-mediated AmpC β-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae* isolates in Iran," *Current Microbiology*, vol. 78, no. 6, pp. 2317–2323, 2021.

[8] D. Ortega-Paredes, P. Barba, S. Mena-López, N. Espinel, V. Crespo, and J. Zurita, "High quantities of multidrug-resistant *Escherichia coli* are present in the Machángara urban river in Quito, Ecuador," *Journal of Water and Health*, vol. 18, no. 1, pp. 67–76, 2020.

[9] I. N. Nwafia, M. E. Ohanu, S. O. Ebede, and U. C. Ozumba, "Molecular detection and antibiotic resistance pattern of extended-spectrum beta-lactamase producing *Escherichia coli* in a Tertiary Hospital in Enugu, Nigeria," *Annals of Clinical Microbiology and Antimicrobials*, vol. 18, no. 1, pp. 41–47, 2019.

[10] T. V. Dalmolin, P. L. Wink, D. de Lima-Morales, and A. L. Barth, "Low prevalence of the mcr-1 gene among carbapenemase-producing clinical isolates of Enterobacterales," *Infection Control & Hospital Epidemiology*, vol. 40, no. 2, pp. 263-264, 2019.

[11] L. M. Avery, C. A. Sutherland, and D. P. Nicolau, "Prevalence of in vitro synergistic antibiotic interaction between fosfomycin and nonsusceptible antimicrobials in carbapenem-resistant *Pseudomonas aeruginosa*," *Journal of Medical Microbiology*, vol. 68, no. 6, pp. 893–897, 2019.

[12] N. Adabara, N. Bakinde, S. Enejiyon, T. Salami, and D. Iorzua, "Detection of extended spectrum beta-lactamase producing *Escherichia coli* from urinary tract infection in general hospital, Minna," *Tanzania Journal of Science*, vol. 46, no. 3, pp. 613–619, 2020.

[13] L. Xiong, P. Hu, and H. Wang, "Establishment of epidemic early warning index system and optimization of infectious disease model: analysis on monitoring data of public health emergencies," *International Journal of Disaster Risk Reduction*, vol. 9, no. 2, pp. 102–125, 2021.

[14] W. Zhang, "Geological disaster monitoring and early warning system based on big data analysis," *Arabian Journal of Geosciences*, vol. 9, no. 1, pp. 131–136, 2020.

[15] Y. Wenying, Y. Jinxia, W. Xin, and S. Min, "Application of early warning nursing system during COVID-19 epidemic in children's hospital," *Nano LIFE*, vol. 11, no. 3, Article ID 2140004, 2021.