An efficient high-throughput screening of high gentamicin-producing mutants based on titer determination using an integrated computer-aided vision technology and machine learning

Xiaofeng Zhu[1], Congcong Du[1], Ali Mohsin[1], Qian Yin[2], Feng Xu[1], Zebo Liu[1], Zejian Wang[1], Ying-Ping Zhuang[1], Ju Chu[1], Xiwei Tian[1], and Mei-Jin Guo[1]

[1]East China University of Science and Technology State Key Laboratory of Bioreactor Engineering
[2]South-Central Minzu University Library

May 9, 2022

Abstract

The 'design-build-test-learn' (DBTL) cycle has been adopted in rational high-throughput screening for obtaining high-yield industrial strains. However, the mismatch between build and test slows the DBTL cycle due to the lack of high-throughput analytical technologies. In this study, a highly-efficient, accurate, and non-invasive detection method of gentamicin (GM) was developed, which can provide timely feedback for the high-throughput screening of high-yield strains. Firstly, a self-made tool was established to obtain datasets in 24-well based on the coloring of cells. Subsequently, the random forest (RF) algorithm was found to have the highest prediction accuracy with 98.5% for the training and 91.3% for verification. Finally, a stable genetic high-yield strain (998U/mL) was successfully screened out in 3005 mutants, which was verified to improve the titer by 72.7% in a 5 L bioreactor. Moreover, the verified new datasets were updated to the model database in order to improve learning ability of DBTL cycle.

Title

An efficient high-throughput screening of high gentamicin-producing mutants based on titer determination using an integrated computer-aided vision technology and machine learning

Authors

Xiaofeng Zhu[1,2], Congcong Du[1,2], Ali Mohisn[1,2], Qian Yin[3], Feng Xu[1,2], Zebo Liu[1,2], Zejian Wang[1,2], Yingping Zhuang[1,2], Ju Chu[1,2], Xiwei Tian[1,2]*, Meijin Guo[1,2]*

Affiliations

[1] State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, 130 Meilong Rd., Shanghai 200237, China

[2] School of Biotechnology, East China University of Science and Technology, 130 Meilong Rd., Shanghai 200237, China

[3] College of Biological & Medical Engineering, South-Central University for Nationalities, Minzu Road 182, Wuhan, Hubei 430070, China

***Corresponding authors**

Prof. Meijin Guo

Address: P.O. box 329#, East China University of Science and Technology, 130 Meilong Rd., Shanghai 200237, P. R. China. Tel: +86 21 64253011; Fax: +86 21 64253702. E-mail:*guo_ mj@ecust.edu.cn***ORCID ID:** 0000-0002-3171-4802

Dr. Xiwei Tian

Address: East China University of Science and Technology, 130 Meilong Rd., Shanghai 200237, P. R. China. Tel: +8613482502988; E-mail:*xiweitian@ecust.edu.cn*

Abstract

The 'design-build-test-learn' (DBTL) cycle has been adopted in rational high-throughput screening for obtaining high-yield industrial strains. However, the mismatch between build and test slows the DBTL cycle due to the lack of high-throughput analytical technologies. In this study, a highly-efficient, accurate, and non-invasive detection method of gentamicin (GM) was developed, which can provide timely feedback for the high-throughput screening of high-yield strains. Firstly, a self-made tool was established to obtain datasets in 24-well based on the coloring of cells. Subsequently, the random forest (RF) algorithm was found to have the highest prediction accuracy with 98.5% for the training and 91.3% for verification. Finally, a stable genetic high-yield strain (998U/mL) was successfully screened out in 3005 mutants, which was verified to improve the titer by 72.7% in a 5 L bioreactor. Moreover, the verified new datasets were updated to the model database in order to improve learning ability of DBTL cycle.

Keywords: Non-invasive method; Gentamicin; Machine learning; High-throughput screening; Computer-aided vision technology.

Introduction

Gentamicins (GMs) belong to one kind of aminoglycoside antibiotics produced by the microorganisms such as *Micromonospora purpura* and*Micromonospora echinospora* (Houghton, Green, Chen, & Garneau-Tsodikova, 2010). It was discovered in 1963 and then introduced into clinical usage in 1971 (C. Chen, Chen, Wu, & Chen, 2014). Because of the low price, wide antibacterial spectrum, strong antibacterial effect, and stable efficacy, it has been widely used in clinics (Liu et al., 2018; Nordang & Anniko, 2005; Sohail, Esquer Garrigos, Elayi, Xiang, & Catanzaro, 2020). Moreover, it is a well-known kind of antibiotic with a large production scale at present due to mature fermentation technology (Liu et al., 2018). However, a higher GM producing strain is considered to be a key factor in the fermentation process which significantly improve the competitiveness of companies (Zhou, Tian, Lin, Zhang, & Chu, 2019). Therefore, the 'design-build-test-learn' (DBTL) cycle is an engineering paradigm that is widely applied in rational high-throughput screening for obtaining high-yield industrial strains. In general, it is easy to create a large mutant library. For example, mutagenesis and genetic engineering, as well as adaptive evolution, are the main approaches to improve the performance of strains to build large libraries of mutants (Zhou et al., 2019). Moreover, the atmospheric and room temperature plasma (ARTP) is a novel mutagenesis technology widely used for microbial strain improvement (Y. Chen et al., 2020; Shu et al., 2020; Wang et al., 2019). However, the rapid and efficient acquisition of target strains from a large strain variants library through high-throughput screening technology remains a great challenge. The main bottleneck technology is low efficient high-throughput analytical methods during DBTL cycle.

Currently, there are several methods for GM detection, including chromatography (high-performance liquid chromatography, HPLC), spectroscopy (nephelometry), and biological method (Aunon et al., 2020; Doadrio et al., 2004; Lukáč et al., 2019). Nevertheless, GM has to be pre-column derivatized for titer detection by HPLC due to the absence of conjugated double bonds in the molecular structure (no UV absorption peaks) (Cabanes et al., 1991). Although this method is precise, the sample stability after derivatization is poor, and it is inefficient for high-throughput screening in rapid detection during the whole mutation breeding process. On the other hand, spectrophotometry is regarded as a rapid detection method. GM reacts in a solution of sodium phosphotungstate under an acidic environment, and the absorbance at a maximum absorption wavelength correlates positively with GM titer, thereby completing rapid GM detection (Tian et

al., 2019). Although this method improves the efficiency, it still suffers from some disadvantages, such as tedious procedure, poor stability, and restricted detection range. In fact, the above methods do not really run DBTL cycles with high efficiency. Besides that, the endpoint assay resulted in more false positives and lower sample screening efficiency during the high-throughput screening of mutants. Therefore, it is urgent to develop a simple, efficient, self-learning, and non-invasive method for screening mutants after cultivation in 24- well plates to achieve the purpose of rapid and high-throughput screening.

Computer vision technology is all the rage at the moment. It is an advanced technology related to computer graphics, image processing, pattern recognition, and machine learning with rapid, non-destructive, real-time, economic characteristics. Moreover, the target image information is acquired via the imaging system and then transmit to the image processing system, which converts the image (color, texture, brightness, pixel distribution, *etc* .) into digital information, which can be further calculated, processed, and resolved for identifying, detecting and controlling the target object. Besides that, this technology has widened application areas because of its stability, flexibility, accuracy, and so forth (Agarwal, Kumar, Varadwaj, & Tiwari, 2020; Memmolo et al., 2022; Phromphithak, Onsree, & Tippayawong, 2021; Zhu et al., 2021). In addition, advances in artificial intelligence are changing all areas of chemistry,and biological processes at an increasingly rapid pace, without an in-depth understanding of the internal operation of biological systems. The results obtained from the phenomena can better guide biological research(Hesami & Jones, 2020; Nandy et al., 2021).

In this study, the computer-aided vision technology upon integration with machine learning was used for the quick and non-invasive detection of GM titer after cultivation in microplates during the primary screening of mutants of *Micromonospora echinospora* , as shown in Fig. 1 (a). *M. echinospora* 49-92S exhibited a coloring effect in the fermentation broth, and there was a strong correlation with GM titer during 0-6 days of culture. Meanwhile, a picture grab tool was developed for 24-well plates by *Python+OpenCV* , which allows rapid and high-throughput pixel capture of 24 wells and generation of a data matrix of 15 features. Then, three widely used machine learning models were evaluated by 768 training datasets, and one model was determined to be the best fit for this dataset based on 50-fold cross-validation scores. In addition, we used 94 datasets of non-identical biological batches to validate the model and found that the random forest (RF) possessed the best prediction effect in the same batch or different batch. To verify the feasibility of the method for quick screening application, the utilization of the proposed method in this study was executed to complete the primary screening of 3005 mutants. As a result, the high-yielding producers were successfully screened and updated the model database. Thus, this systematically modeled approach can be performed for the real-time monitoring of the changes in GM titer due to the non-invasive method preferable to reduce the probability of false positives and increases DBTL clycle efficiency in mutant screening. Moreover, the high-throughput characteristic makes the GM detection from a single well, endpoint (2D) evolved into timelines, multipoint (face) (3D).

Materials and Methods

Image system

The equipment of the established system is consisting of a camera (FUJIFILM X-A2), a fixed light source, a bracket, and an enclosed space. In the system, it is significant for working well to avoid reflection effects. For all experiments, digital images of 24-well plates were saved and stored in JPG format (pixel size is 3264 *3264). Moreover, the parameter of the camera is F5.6, the speed of the shutter is 0.4 s, ISO is 200, manual focus, and manual white balance.

24-well plate images capture and digital matrix extracting

Accurately, the most crucial step was efficiently locating and cutting the position of each hole digital image in the 24-well plate in order to obtain the picture of color changes in GM fermentation process, and effectively converted the images into numerical information. All images were processed to obtain a piece of digital information by an executable file written by *Python+OpenCV* . Image processing operations include the following steps:

Pre-processing, import the image file path; change the color mode to*GRAY;* and *Media blur* was used to eliminate picture noise.

Image binarization, the function of *OpenCV.threshold* was used to binarize images. The image was converted to a binary image according to the custom threshold by the following function:

$$\mathrm{dst}\,(x,y) = \begin{cases} maxval, & if\ src\,(x,y) > thresh \\ 0, & otherwise \end{cases}$$

src: source image; *dst*: output image; *thresh*: threshold; *maxval*: maximum value in *dst* image.

The acceptable threshold value was in the range of 160-190, which can accurately highlight the spatial position of each 24 holes (distinguish the foreground color and background color), as shown in Fig. 1 (b).

Contour detection, the contour detection function (*cv2.findContours(*image, *mode, method)*) was used to detect the information about contour position.*model*: *cv2.RETR_LIST*, contour no hierarchical relationship.*method*: *cv2.CHAIN_APPROX_SIMPLE*, this method can effectively reduce the amount of calculation by compressing the elements in the horizontal direction, vertical direction and diagonal direction, and only retaining the end coordinates in this direction. The minimum circumscribed rectangle was established by the contour information. The new images were cut and retained the circumscribed rectangle that meets the specifications after setting constraints.

Feature extraction, the quadrant of 500 * 500 pixels in the new image was used to calculate the mean of red (R), green (G), blue (B), brightness (L , the mean of R, G and$B$), three relative colors (relative red ($rr = R/L$), relative green ($rg = G/L$), relative blue ($rb = B/L$)), hue (H), saturation (S), *L-lab* (brightness in*Lab* mode), cyan (c), magenta (m), yellow (y), black (k) and fermentation days (day).

Machine learning model selection

The titer of GM was quickly detected by colorimetry, and the images of 8 groups of 24 well plates were obtained at different fermentation stages (4 detection repetitions in each group). The regression model was established by a 768 * 15 matrix. The regression methods include Partial least squares regression (PLS), support vector machine (kernel function: "linear", "RBF", "poly"), RF and gradient lifting decision tree (GBR), using the scikit-learn library in a python programming language*(Python 3.8)* .

The model scores were evaluated by modeling 768 data with 50-fold cross-validation under the default parameters. Then the learning curves were drawn to evaluate the acceptability of models to different amounts of data, and the prediction ability of samples in the same batch (10-fold cross-validation).

Stability test

Six levels of GM titers were used to capture images at 8 time points within 24 hours in *M. echinospora* 49-92S fermentation systems. The principal component analysis (PCA) and Duncan multiple comparison ($P<0.05$) were used to analyze the stability and accuracy of the system in different time periods (Gonzalez-Freire et al., 2018). The closer a data point from the others, the stronger will be the correlation in the PCA scores plot.

RF parameter optimization

RF regression is efficient on large datasets, in which it combines the advantages of predictions from multiple decision tree algorithms (Phromphithak et al., 2021). The *scikit-learn* library module performs RF parameter adjustment. The model function is as follows:*RandomForestClassifie* r *(n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, max_leaf_nodes, random_state =20)* . Parameter adjustment priority:

n_estimators > max_depth > min_samples_leaf > min_samples_split > max_features.

Model validation

GM fermentation was carried out in 24-well plates of 4 groups of non-training dataset batches. Images of different fermentation degrees were captured within 0-6 days, 94 * 15 training dataset matrix was generated for model verification, and the generalization ability of the training model in different biological batches was evaluated.

Determination of GM content (UV spectrophotometry)

First of all, an appropriate amount of fermentation broth was taken and adjusted the pH to 1.5-2.0 with 20% H_2SO_4 solution. Then, the pH was adjusted to 6.4-6.8 via NaOH after 30 minutes of ultrasound sonication. After that, the supernatant was taken for standby after centrifugation at 4000 r/min for 20 minutes. The supernatant of acidified fermentation broth was mixed with the 1.5 mL sodium phosphotungstate aqueous solution. After 30 min, the volume was fixed to 100 mL and centrifuged at 4000 r/min for 10 min. The supernatant was then taken to measure the GM content by using the spectrometer at a wavelength of 450 nm. Finally, GM content was calculated according to the standard curve.

Determination of GM content (HPLC)

Derivatizing agent: phthalaldehyde (1 g) and boric acid (2 g) were dissolved in 100 mL of 5% methanol solution, respectively. After that, 2 mL mercaptoacetic acid was added to this solution. The pH was adjusted to 10.4 with 45% NaOH and then stored in a 4 dark refrigerator.

Mobile phase: sodium heptanesulfonate (4 g) was dissolved in 1 L methanol acetic-acid solution (225 mL ultrapure water, 730 mL methanol, 45 mL acetic acid).

HPLC conditions: the chromatographic separation of GM was performed using Agilent chromatograph (1260 Infinity, HPLC), equipped with a quat pump, UV detector, and a C-18 unitary column (ZORBAX SB; 5 μm, 4.6×150 mm). The GM was detected at 330 nm and the cycle was set for 8 min, with column temperature at 40 °C, injection of 20 μL, and flow rate of 1.4 mL/min.

Sample treatment: 0.2 mL acidified sample was taken in 1.8 mL ultrapure water with 0.8 mL derivant and 2.2 mL methanol. After that, it was stored in a 60 water bath and incubated for 15 minutes.

Media and culture conditions of microorganism and ARTP mutagenesis

M. echinospora 49-92S was used as the original strain for GM production and it was stored at -80°C in 20% glycerol solution. The basal solid medium contained (in g/L) soluble starch 10, asparagine 0.02, CaCO3 1, $MgSO_4{\cdot}7H_2O$ 0.5, KH2$PO_4$0.3, NaCl 0.5, KNO_3 1, agar 13, bran 13. The solid plate was cultured at 35°C for 8-10 days. The basal seed medium contained (in g/L) soluble starch 10, glucose 1, corn flour 15, soybean powder 10, peptone 2, KNO_3 0.5, CaCO3 5. The seed cultivation was carried out at 250 rpm and 34°C for 60 h. The fermentation medium contained (in g/L) soluble starch 30, glucose 5, corn flour 25, soybean powder 26, peptone 10, KNO_3 0.5, CaCO3 7, (NH4)2SO4 1, CoCl 0.03. The fermentation was carried out at 250 rpm and 34°C for 5-6 days.

The fermentation medium in the 5 L bioreactor was the same as described above (Shanghai Guoqiang Bioengineering Equipment Co., Ltd., China). All the media were sterilized at 115°C for 30 min. The initial working volume of 3 L with an inoculum of 12% was cultured at 34°C for 130 h. Aeration at 0.5 vvm and dissolved oxygen (DO) above 30% of the saturation concentration were maintained by adjusting the agitation in a stepwise manner. pH of 7.6 was maintained by the addition of ammonia solution during the whole process.

For ARTP mutagenesis, 10 μL cell suspension was treated by ARTP with a mutagenesis time of 210 s. Single colonies were obtained by culturing on plates. The seed culture was performed in well plates, with 5 single colonies in each well.

Fermentation parameter determination

Residual sugars were measured by dinitrosalicylic acid method (DNS) method (Lee et al., 2013). Microbial packed mass volume (PMV) was used to characterize the cell concentration (Xia et al., 2009). A certain

volume of fermentation broth V_1 was centrifuged at 3000 rpm 10 min. The volume V_2 of the supernatant was rapidly determined, and the cell concentration was calculated from the formula as given below:

$$\mathrm{PMV}\,(\%) = \frac{V_1 - V_2}{V_1} \times 100\%$$

Results and Discussion

Selection of machine learning model

In this study, the imaging stability of system at different times was assessed in eight-time points within 24 h. Also, six levels of GM were chosen for imaging. PCA analysis and multivariate difference analysis were performed on the sample matrix. As shown in Fig. 2 (a), samples of the same titer maintained a higher aggregation degree at different times and better separation between different titers. Multivariate analysis ($P < 0.01$) of the feature factors were listed in Table 1, where factors R, G, rr, rg, m, and k all showed extremely significant differences in the parameters between different titers, while others also showed different degrees of differences. Thus, these results illustrated that the imaging system had excellent reproducibility in different time dimensions.

Several machine learning models were employed to model the color features of GM from *M. echinospora* 49-92S, including three support vector machines with different kernel functions (SVM_linear, SVM_rbf, and SVM_poly), PLS, RF, and GBR. 50-fold cross-validation of the 768 *15 dataset modeling was performed under default model parameters, and the model scores are shown in Fig. 2 (b), where RF, GBR and PLS are better suited for the dataset than SVM. The score of ensemble algorithm RF and GBR performed were higher than PLS. These results suggest that an ensemble algorithm could provide a higher prediction performance than a single approach (Duran et al., 2018). The error decreases as the number of samples increases by contrasting the learning curve scores (10-fold cross-validated), where PLS, RF and GBR could converge on a higher score level compared to SVM, especially RF has a higher training set score (98.2%). Furthermore, the results showed that training RF from the whole dataset provided the highest performance model in the same biological batch, as shown in Fig. 2 (c-h).

RF regulation and model validation

In machine learning, overfitting and underfitting models will lead to the increase in generalization error (Eken, 2021; Salam, Azar, Elgendy, & Fouad, 2021). RF could be balanced by adjusting parameters and the prediction power for unknown samples would be fully realized (Hou et al., 2021; Torre-Tojal, Bastarrika, Boyano, Lopez-Guede, & Graña, 2022). 768 sample datasets were modeled under the default parameters. Whereas, 94 sample data from different batches were used as the validation set for model validation. The scores of RF and GBR training sets were $R^2 = 0.982$ and $R^2 = 0.976$ respectively, and the verification sets were $R^2 = 0.893$ and $R^2 = 0.887$. The score of PLS training set $R^2 = 0.913$, verification set $R^2 = 0.881$, as shown in the following Fig. 3 (a, b, c).

The parameter *n_estimators,max_depth,min_samples_leaf, max_features* were adjusted by plotting the learning curve. Under the default parameters, the model score $R^2 = 0.9821$. The optimal parameter *n_estimators = 186* and the model score is 0.98264. The complexity of the model was unchanged. The optimal *max_depth = 19* caused the score to 0.98265, and the complexity moved to a simple direction. The parameters of *min_samples_leaf* were adjusted for continuing to move in the simple direction, and the score decreases. The model in the simple direction has moved to the limit. The *max_features* the larger, the more complex the model was (default was 3). The maximum score was 0.98469 *(features = 6)* . At this time, the model had reached the parameter adjustment limit, the score of regression was about 98.469% for the training set and 91.3% for verification. After parameter adjustment, the score of the model training set was increased by 0.259%, and the score of the verification set is increased by 2%, as shown in Fig. 3 (d).

Feature factor

Heat maps were drawn for 15 feature factors and titers, as shown in the following Fig. 3 (e). There was an obvious correlation between the factors change trend involved in the modeling and the GM titer of*M. echinospora* 49-92S. Through *Pearson* correlation coefficient analysis, the factors with strong correlation (*|r|[?]0.8*) were r, G, B, L, k. While, the factors of moderate correlation (*0.5[?]|r|<0.8*) were H, S, RR, l-lab, c day, and the factors of low correlation (*0.3[?]|r|<0.5*) were rg, rb, m, y. There were no unrelated factors were found (*|r|<0.3*). The importance of the characteristic factor was returned via function *feature_importance* , as shown in Fig. 3 (f), in which the factors of L, R, and k (*Pearson |r|[?]0.8*) also showed a strong contribution to the model. In addition, the "day" variables as environmental parameters are also incorporated into the machine learning model to improve the accuracy and robustness. Removing the features of rg, m, c, etc. with low contribution reduces the score, indicating that the selection of 15 groups of feature factors in this model was a necessary condition to maintain the high quality of the model.

Comparison of detection methods

GM is a group of multi-component aminoglycoside antibiotics with similar structures produced by *M. echinospora* 49-92S. Therefore, it is divided into group C and group A/B according to the difference in molecular structure, in which component C is regarded as an effective component (Wagman, Oden, & Weinstein, 1968). In this study, it was necessary to accurately quantify the samples by HPLC after derivatization treatment, and HPLC of the derivatized standards mainly contains four peaks, C_1, C_{1a}, C_2, C_{2A} as shown in the following Fig. 4 (a). The linear relationship between GM titer and peak area was established by averaging the areas of four peaks, $R^2 = 0.999$. Spectrophotometry was more efficient than HPLC in the rapid screening of GM mutagenesis because of without derivation (Kumar, Himabindu, & Jetty, 2008; Tian et al., 2019). Its accuracy has also been experimentally confirmed in this study, and an excellent linear relationship $R^2 = 0.99$, as shown in Fig. 4 (b). However, this method must terminate the fermentation of microporous culture, and the complex pretreatment process also makes its accuracy affected by the proficiency of operators such as acidification, matching measuring range, etc. As a result, the current detection efficiency was still not matched with a large number of mutation samples, the cycle of DBTL was broken between build and test.

In this work, firstly, the tool has completed the upgrading and promotion than the previous screening works, as presented in Table 2. The 862 high coverage GM fermentation samples with different titers ensure the same high accuracy of the model as spectrophotometry and HPLC in the same batch ($R2 > 0.99$). The positive sample after verification could improve the quality and robustness of model to avoid detection error due to the exceeding the range in the above method.

Besides that, different from (Zhu et al., 2021), who combined computer vision technology with machine learning applied in plant cell culture, which realized the breakthrough of plant cell culture from point to a line in shake flask (1D to 2D). In this work, there is an advancement from a single point (1D) evolved into the surfaces (3D) for quick screening technology. Therefore, it could be rapid, accurate, and non-invasive detection for high-through screening, which ensured the precision. For example, the detection efficiency was 3.83-fold higher than for spectrophotometry and 228-fold higher than for HPLC because of the above advantages.

Moreover, it could realize the real-time high-throughput process detection of GM synthesis without stopping the fermentation process, as shown in Fig. 4 (c). Now the samples can be screened by a multi-dimensional scheme. For example, the total yield, specific yield, product synthesis rate, and other process parameters can be considered comprehensively to screen the mutants, which is an advantage that endpoint detection does not execute. This simplifies the tedious operation, reduces the probability of false positives, and maximizes the efficiency of rapid screening.

High-throughput screening of high-yield GM mutants

The method developed in this work could realize rapid, real-time, non-invasive and high-throughput detection of GM titer in 24-well plate fermentation culture. 3005 mutants were obtained by ARTP mutagenesis. After 24-well plate culture for preliminary screening, the determination results were shown by this method in the Fig. 5 (a). Among them, 405 mutant strains present higher titers than that of the parent strain, of which

the highest titer is 988.0 U/mL, 71.0% higher than the control (577.9 U/mL). Subsequently, 175 mutants (10% higher than the control group) were re-screened by shake flask, as shown in the Fig. 5 (b). The 15 mutants above 10% of the control group were obtained. Among them, the highest titer of re-screening was 998 U/mL, which still exhibited high-yielding characteristics after 5 consecutive generations of plate transfer culture.

Significant differences in GM titers during fed-batch fermentation were observed between mutant and parent strains when the same inoculum size, ventilation, temperature, and rotation rate of the fermentation phase were maintained. The mutant strain presented an immense advantage in GM titer in contrast to the parent strain in the whole fermentation process, as shown in Fig. 5 (c). Moreover, the mutant showed a faster rate of growth and carbon source consumption Fig. 5 (d, e). Thus, all these results indicated that the mutant strain had a stronger metabolic capacity and production level.

Future prospects

The rapid, accurate, and process features make the final titer the only screening index in the high-throughput screening. This would cause some 'seed players' to be ignored, which have outstanding production potential only in the upfront period because culture conditions limit total yield. Following up on work we hope to introduce more features besides color into the model to achieve comprehensive detection of biomass, titer, etc. In addition, the process features such as specific productivity, biomass, and total production as a comprehensive index to establish multidimensional screening network. Therefore, the screening method developed here could provide researchers with a novel and efficient way for high-yield antibiotic-producing mutants selection. In addition, the DBTL cycle will be adopted in rational high-throughput screening for high-yield industrial strains in quality, fast, and intellect patterns, as shown in Fig. 6.

Conclusions

In this work, the phenomenon of coloring in GM fermentation was a real-time photo captured in 24-well plates by a self-made laboratory image recognition tool, establishing the model database. RF model was successfully applied to predict the titer of GM, using highly influential 15-features consisting of color and time characteristics. The score of model was about 98.5% for the training and 91.3% for verification. In addition, the process titer of 3005 mutants was screened by this method, and a high-yielding strain (998 U/ml) was successfully observed, which was verified to improve the titer by 72.7% in a 5 L bioreactor.

Acknowledgments

This work was financially supported by the National Key Research Development Program of China (No. 2019YFA0904800).

Conflicts of interest

There are no conflicts to declare.

Data availability statement

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

References

Agarwal, V., Kumar, D., Varadwaj, P., & Tiwari, A. (2020). Water activity and biomass estimation using digital image processing in solid-state fermentation. *Bioresour Technol, 308* , 123277. doi:10.1016/j.biortech.2020.123277

Aunon, A., Esteban, J., Doadrio, A. L., Boiza-Sanchez, M., Mediero, A., Eguibar-Blazquez, D., . . . Aguilera-Correa, J. J. (2020). Staphylococcus aureus Prosthetic Joint Infection Is Prevented by a Fluorine- and Phosphorus-Doped Nanostructured Ti-6Al-4V Alloy Loaded With Gentamicin and Vancomycin. *J Orthop Res, 38* (3), 588-597. doi:10.1002/jor.24496

Cabanes, A., Cajal, Y., Haro, I., Anton, J. M. G., Reig, F., & Arboix, M. (1991). Gentamicin Determination in Biological Fluids by HPLC, Using Tobramycin as Internal Standard. *Journal of Liquid Chromatography, 14* (10), 1989-2010. doi:10.1080/01483919108049669

Chen, C., Chen, Y., Wu, P., & Chen, B. (2014). Update on new medicinal applications of gentamicin: evidence-based review. *J Formos Med Assoc, 113* (2), 72-82. doi:10.1016/j.jfma.2013.10.002

Chen, Y., Tian, X., Li, Q., Li, Y., Chu, J., Hang, H., & Zhuang, Y. (2020). Target-site directed rational high-throughput screening system for high sophorolipids production by Candida bombicola. *Bioresour Technol, 315* , 123856. doi:10.1016/j.biortech.2020.123856

Doadrio, A. L., Sousa, E. M., Doadrio, J. C., Perez Pariente, J., Izquierdo-Barba, I., & Vallet-Regi, M. (2004). Mesoporous SBA-15 HPLC evaluation for controlled gentamicin drug delivery. *J Control Release, 97* (1), 125-132. doi:10.1016/j.jconrel.2004.03.005

Duran, C., Daminelli, S., Thomas, J. M., Haupt, V. J., Schroeder, M., & Cannistraci, C. V. (2018). Pioneering topological methods for network-based drug-target prediction by exploiting a brain-network self-organization theory. *Brief Bioinform, 19* (6), 1183-1202. doi:10.1093/bib/bbx041

Eken, E. (2021). Determining overfitting and underfitting in generative adversarial networks using Frechet distance. *Turkish Journal of Electrical Engineering & Computer Sciences, 29* (3).

Gonzalez-Freire, M., Scalzo, P., D'Agostino, J., Moore, Z. A., Diaz-Ruiz, A., Fabbri, E., . . . Ferrucci, L. (2018). Skeletal muscle ex vivo mitochondrial respiration parallels decline in vivo oxidative capacity, cardiorespiratory fitness, and muscle strength: The Baltimore Longitudinal Study of Aging. *Aging Cell, 17* (2). doi:10.1111/acel.12725

Hesami, M., & Jones, A. M. P. (2020). Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. *Appl Microbiol Biotechnol, 104* (22), 9449-9485. doi:10.1007/s00253-020-10888-2

Hou, H., Zhu, S., Geng, H., Li, M., Xie, Y., Zhu, L., & Huang, Y. (2021). Spatial distribution assessment of power outage under typhoon disasters. *International Journal of Electrical Power & Energy Systems, 132* , 107169.

Houghton, J. L., Green, K. D., Chen, W., & Garneau-Tsodikova, S. (2010). The future of aminoglycosides: the end or renaissance?*Chembiochem, 11* (7), 880-902. doi:10.1002/cbic.200900779

Kumar, C., Himabindu, M., & Jetty, A. (2008). Microbial Biosynthesis and Applications of Gentamicin: A Critical Appraisal. *Critical Reviews in Biotechnology, 28* (3), 173-212. doi:10.1080/07388550802262197

Lee, O. K., Kim, A. L., Seong, D. H., Lee, C. G., Jung, Y. T., Lee, J. W., & Lee, E. Y. (2013). Chemoenzymatic saccharification and bioethanol fermentation of lipid-extracted residual biomass of the microalga, Dunaliella tertiolecta. *Bioresour Technol, 132* , 197-201. doi:10.1016/j.biortech.2013.01.007

Liu, Y., Feng, Y., Cheng, D., Xue, J., Wakelin, S., & Li, Z. (2018). Dynamics of bacterial composition and the fate of antibiotic resistance genes and mobile genetic elements during the co-composting with gentamicin fermentation residue and lovastatin fermentation residue.*Bioresour Technol, 261* , 249-256. doi:10.1016/j.biortech.2018.04.008

Lukač, P., Hartinger, J., Mlček, M., Popkova, M., Suchy, T., Šupová, M., . . . Grus, T. (2019). A novel gentamicin-releasing wound dressing prepared from freshwater fish Cyprinus carpio collagen cross-linked with carbodiimide. *Journal of Bioactive and Compatible Polymers, 34* , 088391151983514. doi:10.1177/0883911519835143

Memmolo, P., Aprea, G., Bianco, V., Russo, R., Andolfo, I., Mugnano, M., . . . Ferraro, P. (2022). Differential diagnosis of hereditary anemias from a fraction of blood drop by digital holography and hierarchical machine learning. *Biosens Bioelectron, 201* , 113945. doi:10.1016/j.bios.2021.113945

Nandy, A., Duan, C., Taylor, M. G., Liu, F., Steeves, A. H., & Kulik, H. J. (2021). Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chem Rev, 121* (16), 9927-10000. doi:10.1021/acs.chemrev.1c00347

Nordang, L., & Anniko, M. (2005). Nitro-L-arginine methyl ester: A potential protector against gentamicin ototoxicity. *Acta Oto-Laryngologica, 125* (10), 1033-1038. doi:10.1080/00016480510038022

Phromphithak, S., Onsree, T., & Tippayawong, N. (2021). Machine learning prediction of cellulose-rich materials from biomass pretreatment with ionic liquid solvents. *Bioresour Technol, 323* , 124642. doi:10.1016/j.biortech.2020.124642

Salam, M. A., Azar, A. T., Elgendy, M. S., & Fouad, K. M. (2021). The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem. *Int. J. Adv. Comput. Sci. Appl., 12* (4), 641-655.

Shu, L., Si, X., Yang, X., Ma, W., Sun, J., Zhang, J., . . . Gao, Q. (2020). Enhancement of Acid Protease Activity of Aspergillus oryzae Using Atmospheric and Room Temperature Plasma. *Front Microbiol, 11* , 1418. doi:10.3389/fmicb.2020.01418

Sohail, M. R., Esquer Garrigos, Z., Elayi, C. S., Xiang, K., & Catanzaro, J. N. (2020). Preclinical evaluation of efficacy and pharmacokinetics of gentamicin containing extracellular-matrix envelope.*Pacing Clin Electrophysiol, 43* (3), 341-349. doi:10.1111/pace.13888

Tian, J. T., Li, M. C., Hang, H. F., Guo, M. J., Chu, J., & Zhuang, Y. P. (2019). UV spectrophotometric determination of the concentration of gentamicin C_(1a) in the application of high throughput screening.*Chinese Journal of Antibiotics* .

Torre-Tojal, L., Bastarrika, A., Boyano, A., Lopez-Guede, J. M., & Graña, M. (2022). Above-ground biomass estimation from LiDAR data using random forest algorithms. *Journal of Computational Science, 58* , 101517.

Wagman, G. H., Oden, E. M., & Weinstein, M. J. (1968). Differential chromatographic bioassay for the gentamicin complex. *Applied microbiology, 16* (4), 624-627.

Wang, J., Liu, F., Su, T., Chang, Y., Guo, Q., Wang, Q., . . . Qi, Q. (2019). The phage T4 DNA ligase in vivo improves the survival-coupled bacterial mutagenesis. *Microb Cell Fact, 18* (1), 107. doi:10.1186/s12934-019-1160-7

Xia, J.-Y., Wang, Y.-H., Zhang, S.-L., Chen, N., Yin, P., Zhuang, Y.-P., & Chu, J. (2009). Fluid dynamics investigation of variant impeller combinations by simulation and fermentation experiment.*Biochemical Engineering Journal, 43* (3), 252-260. doi:10.1016/j.bej.2008.10.010

Zhou, G., Tian, X., Lin, Y., Zhang, S., & Chu, J. (2019). Rational high-throughput system for screening of high sophorolipids-producing strains of Candida bombicola. *Bioprocess Biosyst Eng, 42* (4), 575-582. doi:10.1007/s00449-018-02062-w

Zhu, X., Mohsin, A., Zaman, W. Q., Liu, Z., Wang, Z., Yu, Z., . . . Chu, J. (2021). Development of a novel noninvasive quantitative method to monitor Siraitia grosvenorii cell growth and browning degree using an integrated computer-aided vision technology and machine learning.*Biotechnol Bioeng, 118* (10), 4092-4104. doi:10.1002/bit.27886

Tables

Table 1

The mean, SD, and MA about 15 features in six levels GM titers at 8-time points.

Parameter	Titer	Mean	SD	MA	Parameter	Mean	SD	MA	Parameter	Mean	SD	M
R	326	0.55948	0.02405	A	S	0.77459	0.02670	B	L(Lab)	0.07307	0.00894	B

Parameter	Titer	Mean	SD	MA	Parameter	Mean	SD	MA	Parameter	Mean	SD	M
	392	0.49601	0.01189	B		0.77782	0.01656	B		0.05619	0.00070	BC
	450	0.45512	0.01146	C		0.80807	0.01116	A		0.06613	0.00045	BC
	586	0.16287	0.01413	D		0.59727	0.00971	C		0.02562	0.00137	C
	790	0.11471	0.00964	E		0.47874	0.02516	D		0.06623	0.03706	BC
	863	0.07912	0.00691	F		0.43262	0.01213	E		0.49438	0.06056	A
G	326	0.45965	0.02397	A	L	0.39630	0.02166	A	c	0.00000	0.00000	B
	392	0.30730	0.01327	C		0.30552	0.01193	B		0.00000	0.00000	B
	450	0.35059	0.01130	B		0.30166	0.01007	B		0.00000	0.00000	B
	586	0.10695	0.00950	E		0.11254	0.01019	C		0.00000	0.00000	B
	790	0.19259	0.01021	D		0.12567	0.00920	C		0.39800	0.02251	A
	863	0.06984	0.00541	F		0.06614	0.00588	D		0.00000	0.00000	B
B	326	0.14669	0.02208	A	rr	0.15942	0.00621	C	m	0.16795	0.00936	D
	392	0.11326	0.01102	B		0.18048	0.00295	A		0.38067	0.01321	A
	450	0.09925	0.00791	B		0.16768	0.00158	B		0.22979	0.00647	C
	586	0.06779	0.00763	C		0.16087	0.00236	C		0.34295	0.02583	B
	790	0.06892	0.00770	C		0.10221	0.00144	E		0.00000	0.00000	F
	863	0.04946	0.00541	D		0.13293	0.00110	D		0.11643	0.01756	E
H	326	0.11200	0.00108	B	rg	0.13785	0.00720	B	y	0.75332	0.02870	B
	392	0.07960	0.00099	BC		0.11174	0.00060	E		0.77198	0.01713	AI
	450	0.09368	0.00064	BC		0.12914	0.00034	C		0.78215	0.01270	A
	586	0.03629	0.00193	C		0.10565	0.00288	F		0.58458	0.01197	D
	790	0.09386	0.05248	BC		0.17020	0.00388	A		0.64565	0.02149	C
	863	0.70037	0.08579	A		0.11744	0.00181	D		0.37587	0.01541	E
S	326	0.77459	0.02670	B	rb	0.04094	0.00406	D	k	0.43910	0.02492	F
	392	0.77782	0.01656	B		0.04111	0.00242	D		0.50399	0.01189	E
	450	0.80807	0.01116	A		0.03651	0.00178	E		0.54488	0.01146	D
	586	0.59727	0.00971	C		0.06682	0.00164	B		0.83713	0.01413	B
	790	0.47874	0.02516	D		0.06075	0.00241	C		0.80702	0.01069	C
	863	0.43262	0.01213	E		0.08296	0.00175	A		0.92088	0.00691	A

Abbreviation: GM, gentamicin; SD, standard deviation; MA, multivariate analysis ($P < 0.01$).

Table 2

Comparison of GM detection methods

	Preprocessing	Derivatization	Continuity	Precision (R^2)	Destructive	Time (hour)	Num
HPLC	Yes	Yes	No	≥0.99	Yes	2	1
Spectrophotometry	Yes	No	No	≥0.99	Yes	0.042	1
Novel method	No	No	Yes	≥0.99	No	0.0087	24

Note: "Time" indicates spend time (hour) to detect one sample.

Abbreviation: GM, gentamicin.

Figure caption

Fig. 1 (a) High-throughput screening procedure. (b) The binary images of GM fermentation on a 24-well plate at thresholds 110, 130, 160, 180, 190, 200.

Fig. 2 (a) Principal component analysis of 6 GM titers at 8-time points, PC is the principal component

contribution. (b) Model scores of 50-fold cross-validation in machine learning models. (c-h) The scores change of training set and testing set (all from 768 * 15 matrix) with data volume (learning curve), and "linear", "rbf", "poly" are three kernel functions of SVM in (c), (d), (e).

Fig. 3 (a), (b), (c) are the scores of machine learnings in default parameters. (d) is the model score under optimal parameters. The "TRAIN" is the training set matrix 768 * 15, and "Verifi." is the validation set matrix 94 * 15, which is not the same batch as "TRAIN". (e) left is a correlation between GM titer (y_t) and feature factors, and right is a correlation between factors.

Fig. 4 (a) shows liquid chromatogram of eight different titers of GM standard samples. (b) shows a linear relationship between the methods of HPLC and spectrophotometry under different GM titers. (c) shows the change of GM titer and color were measured by the novel method in the 24-well-plate mutant screening.

Fig. 5 (a) Preliminary screening of 3005 mutants was completed by this rapid detection; (b) the titers of 175 mutants were re-screened by rapid detection; note: one point represents five mutants. (c), (d), (e) represent the changes of total carbohydrate, PMV and GM of mutant and parent strain in 5 L bioreactor respectively.

Fig. 6 The design-build-test-learn cycle in high-yield antibiotic-producing mutants selection.

Figures

Fig.1

Hosted file

`image1.emf` available at https://authorea.com/users/481354/articles/568448-an-efficient-high-throughput-screening-of-high-gentamicin-producing-mutants-based-on-titer-determination-using-an-integrated-computer-aided-vision-technology-and-machine-learning

Fig. 2

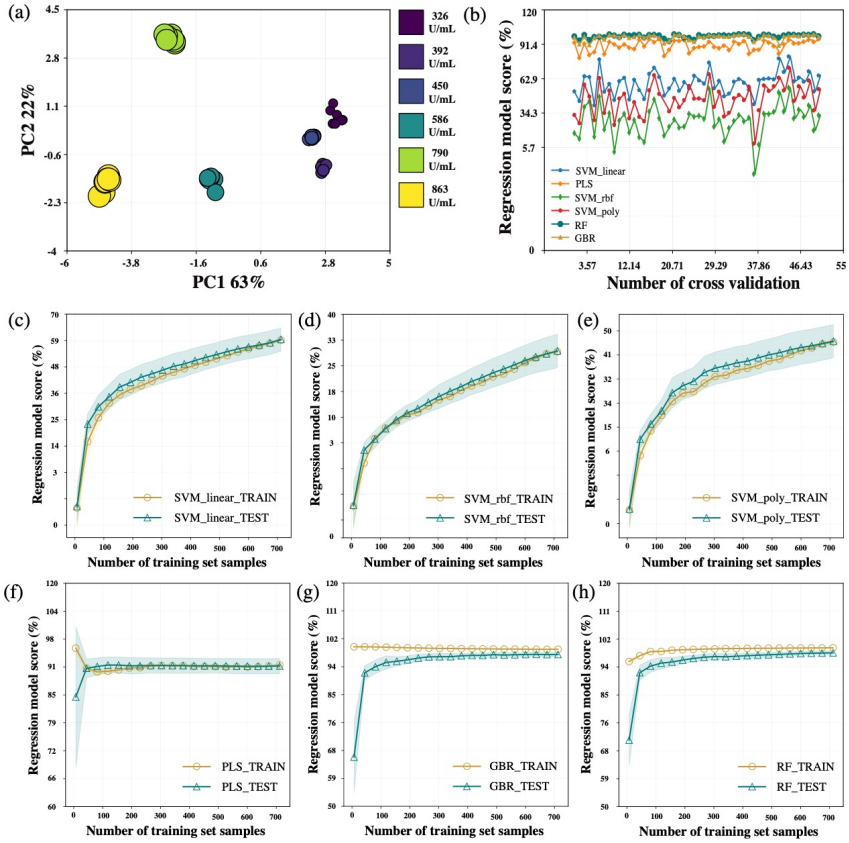

Fig. 3

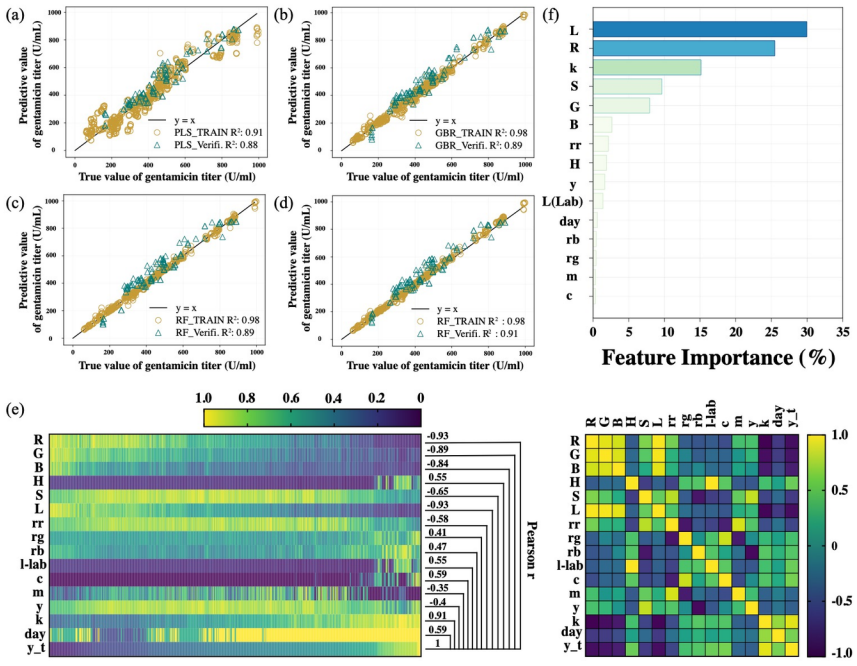

Fig. 4

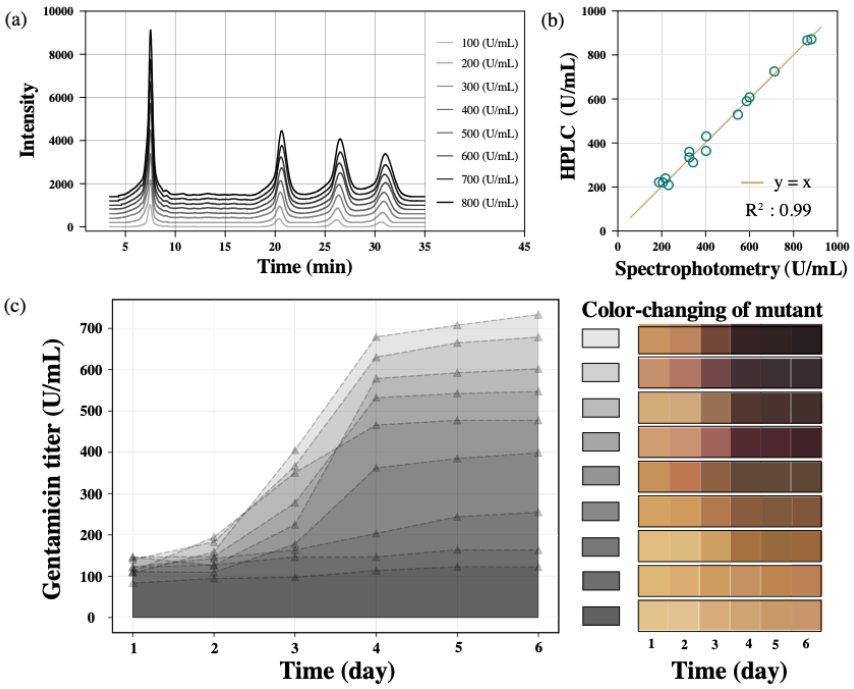

Fig. 5

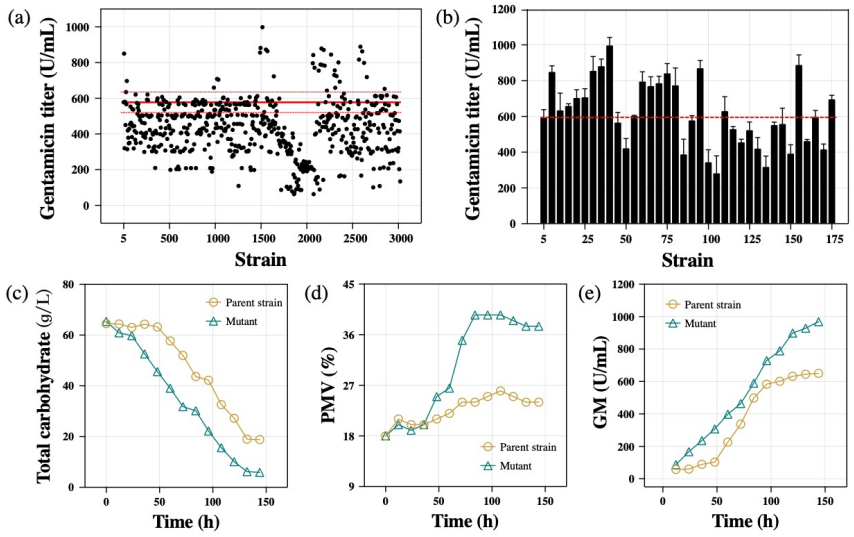

Fig.6